AD-A210 435

DTIC ACCESSION NUMBER

PHOTOGRAPH THIS SHEET

LEVEL

1

INVENTORY

RID 6010-MA-02

DOCUMENT IDENTIFICATION

14 JUNE 1988

DAJA45-88-M-0166

This document has been approved for public release and sale; its distribution is unlimited.

DISTRIBUTION STATEMENT

ACCESSION FOR

NTIS GRA&I ☒

DTIC TAB □

UNANNOUNCED □

JUSTIFICATION

BY

DISTRIBUTION /

AVAILABILITY CODES

DIST	AVAIL AND/OR SPECIAL	
A-1		

DISTRIBUTION STAMP

DTIC COPY INSPECTED 2

DTIC
ELECTE
JUL 21 1989
S E D

DATE ACCESSIONED

DATE RETURNED

89 7 12 009

DATE RECEIVED IN DTIC

REGISTERED OR CERTIFIED NO.

PHOTOGRAPH THIS SHEET AND RETURN TO DTIC-FDAC

DTIC FORM 70A
MAR 86

DOCUMENT PROCESSING SHEET

PREVIOUS EDITION MAY BE USED UNTIL STOCK IS EXHAUSTED.

TROISIEME
WORKSHOP
INTERNATIONAL

THIRD
INTERNATIONAL
WORKSHOP

BELLMAN CONTINUUM

Co-organisé par / Co-organized by
INRIA - UNIVERSITE PARIS 7

13 - 14 JUIN, 1988
JUNE 13 - 14, 1988

SOPHIA - ANTIPOLIS

FRANCE

Actes provisoires
Preprints

⇒ Partial

Table des matières / Contents

COMMANDE DES SYSTEMES DYNAMIQUES INCERTAINS
CONTROL OF UNCERTAIN DYNAMICAL SYSTEMS

—> next page

MODELISATION ET COMMANDE EN ECONOMIE
MODELS AND CONTROL POLICIES IN ECONOMICS

A

SYSTEMES STOCHASTIQUES ET QUANTIQUES

STOCHASTIC AND QUANTUM SYSTEMS

MODELISATION ET COMMANDE DES SYSTEMES BIOLOGIQUES ET DES ECOSYSTEMES

MODELS AND CONTROL POLICIES FOR BIOLOGICAL SYSTEMS AND ECOSYSTEMS

MATHEMATIQUES ET SYSTEMES, ASPECT CALCUL

MATHEMATICS AND SYSTEMS, COMPUTATIONAL BEARINGS

France

C

Troisième
Workshop
International

Third
International
Workshop

BELLMAN

CONTINUUM

Co-organisé par / Co-organized by

INRIA - UNIVERSITE PARIS 7

Patronné par / Sponsored by

AFCET Association Française pour la
Cybernétique Economique et Technique

IFAC International Federation of
Automatic Control
Technical Committee on MATHEMATICS OF CONTROL
Technical Committee on THEORY

Accueilli par / Hosted by

INRIA

SOPHIA - ANTIPOLIS
FRANCE

ACTES PRELIMINAIRES / PREPRINTS

AVANT-PROPOS

Richard Bellman, un des mathématiciens les plus féconds et les plus renommés des Etats Unis, a apporté des contributions majeures aux mathématiques pures et à de nombreux domaines d'applications : sciences de l'ingénieur, économie, médecine, énergie, gestion des ressources en eau, physique mathématique, recherche opérationnelle, sciences de la gestion, psychologie et sociologie. Une telle variété des domaines abordés et des moyens mis en oeuvre pour approfondir ces domaines avec une telle pénétration se rencontre rarement en science.

Tout au long du développement de son oeuvre, il eut un grand nombre d'amis, d'élèves et de correspondants portés vers les mêmes centres d'interêt. Parmi eux, après la disparition du Professeur Bellman, un groupe de scientifiques des Etats Unis s'est efforcé de perpêtuer son Ecole. Dans ce but ils ont proposé d'organiser un Colloque annuel ou bi-annuel : le Bellman Continuum. Ce Colloque devait être de nature interdisciplinaire, comme l'était l'oeuvre de Richard Bellman.

Le premier congrès s'est tenu à l'Université du Michigan, Ann Arbor, Michigan, en 1985 et le second a été accueilli par l'Institut de Technologie de Georgie, Atlanta, Georgie, en 1986. Les organisateurs ont pensé que la France serait un des pays les mieux adaptés à la tenue du troisième congrès pour des raisons de caractère à la fois scientifique et géographique : Richard Bellman était très populaire en Europe. De plus, un argument important pour ce choix était le fait que la huitième Conférence Internationale Analyse et Optimisation des Systèmes de l'INRIA devait se tenir à Antibes du 8 au 10 Juin 1988. Celà fournissait l'occasion idéale de profiter de la présence en un même lieu d'un grand nombre de spécialistes venant de tous les points du monde pour organiser une petite conférence permettant un échange d'idées assez informel.

Dans les deux premiers congrès, le programme avait été dicté par la nature interdisciplinaire du Colloque avec des sujets définis suivant les interêts des participants. Les thèmes unificateurs étaient l'idéologie scientifique et les techniques mathématiques plutôt que les domaines d'étude spécifiques. Dans ce troisième congrès, pour des raisons scientifiques évidentes, l'ensemble des sujets abordés a été délibérément restreint, ayant en vue le fait que ces sujets pourraient changer d'un congrès au suivant. Les sujets mentionnés ci-dessous, choisis dans des domaines où la recherche est très active et pleine de promesses, ont été sélectionnés :

Modélisation et commande en Economie et en Sciences Sociales.
Commande des systèmes dynamiques incertains.
Commande et filtrage nonlinéaire des processus quantiques.
Modélisation et commande des systèmes biologiques.

Les Conférenciers d'Ouverture de Sessions sont les Professeurs

R.E. KALMAN, University of Florida, U.S.A., et Technische Hochschule, Zürich, Suisse
G. LEITMANN, University of California, Berkeley, U.S.A.
S. MITTER, Massachusetts Institute of Technology, U.S.A.

Initialement, notre intention était de réunir un petit nombre de spécialistes sur la base d'invitations. Cependant, les réponses à notre annonce préliminaire surpassèrent notre estimation la plus optimiste de l'enthousiasme des chercheurs dans ces domaines. Par la suite, nous décidâmes d'éditer les Actes de ce Colloque sous la forme d'un livre réunissant les conférences sur invitation et certains des rapports destinés à la présentation de travaux récents, soumis au Comité d'Organisation. Ce livre sera publié après le congrès par SPRINGER-VERLAG dans la Série "Lecture Notes in Control and Information Sciences". Les manuscrits contenus dans le présent fascicule sont les résumés ou les textes intégraux de tous les papiers en notre possession au moment du congrès. Pour chaque thème, dans toute la mesure du possible, ils sont présentés dans l'ordre où ils se trouvent dans le Programme.

Ce Colloque n'aurait pu avoir lieu sans le soutien technique et financier de l'INRIA à qui nous exprimons notre gratitude. En particulier, qu'il me soit permis de remercier ici son Président le Professeur Alain BENSOUSSAN, le Directeur du Centre de Recherche de l'INRIA-SOPHIA ANTIPOLIS le Professeur Pierre BERNHARD, et le Directeur des Relations Internationales et Industrielles Georges NISSEN. J'adresse personnellement des remerciements tout particuliers à Thérèse BRICHETEAU qui, à la tête du Service des Relations Exterieures de l'INRIA, a pris soin si efficacement de la multitude des problèmes d'organisation, et nous a fait profiter de sa grande experience. Nous sommes très reconnaissants à Catherine JUNCKER qui a pris en main de façon experte l'organisation du congrès à Sophia et son implentation sur le site. Je suis redevable envers les Secrétaires Scientifiques de la huitième Conférence Analyse et Optimisation des Systèmes de l'INRIA de leur aide, notamment envers Frédéric BONNANS pour ses conseils éclairés dans la préparation du Colloque. Des remerciements tout particuliers sont dûs à Gilbert MALLET qui a édité ces "Preprints". Finalement, et non les moindres, nos remerciements s'adressent à toutes les secrétaires et les techniciens qui ont contribué au succès du troisième Bellman Continuum, aux différentes étapes et aux différents niveaux de son organisation.

Ce Colloque international a aussi reçu le soutien financier de l'Université Paris 7, l'un des organisateurs du congrès, de l'Association Française pour la Cybernétique Economique et Technique (AFCET), du Centre National de la Recherche Scientifique (CNRS), de l'European Research Office United States Army, du Ministère des Affaires Etrangères, du Ministère de l'Education Nationale, du Ministère de la Recherche et de l'Enseignement Supérieur, de l'United Nations Educational Scientific and Cultural Organization (UNESCO). D'autres subventions ont été attribuées aux participants par divers organismes mentionnés séparément, auxquels, comme aux organismes dont nous venons de dresser la liste, nous exprimons notre gratitude.

Austin BLAQUIERE (Président)

FOREWORD

Richard Bellman, a most prolific and renowned mathematician of the United States, has made major contributions in pure mathematics and in numerous areas of applications : engineering, economics, medicine, energy, water resources, mathematical physics, operations research, management sciences, psychology and sociology. This breadth of interests and this ability to contribute to so many fields at such a high level is rare indeed.

Throughout his years in science, he had a large number of scientific friends, students and followers. Among them, after Professor Bellman has passed away, a group of scientists of the United States has attempted to preserve his School. As a mechanism for achieving this goal, they suggested an annual or biennial workshop : the Bellman Continuum. This workshop was envisioned as being interdisciplinary in nature, as the achievement of Richard Bellman was.

The first meeting was held at the University of Michigan, Ann Arbor, Michigan, in 1985 and the second was hosted by the Georgia Institute of Technology, Atlanta, Georgia, in 1986. The organizers thought that France could be a nice place for the third meeting from both scientific and geographical points of view : Richard Bellman was very popular in Europe. Also, a strong motivation for this choice was the fact that the eighth International Conference Analysis and Optimization of Systems of INRIA was to be held in Antibes on June 8-10, 1988. It provided an ideal opportunity for taking advantage of the presence of a large number of specialists from all parts of the world to organize a small conference where a free exchange of ideas could take place.

In the two first meetings, the program has been dictated by the interdisciplinary nature of the workshop with topics defined by the interest of the participants. The unifying theme included scientific ideology and mathematical tools rather than specific fields of study. In this third one, for evident scientific purposes the subject matter to be treated has been limited, having in view the fact that the areas defined below could change from one meeting to the next. The following topics, chosen in areas where research is very active and promising, in directions opened and explored by Richard Bellman, have been selected :

Models and Control Policies in Economics and Social Systems.
Control of Uncertain Dynamical Systems.
Control and Nonlinear Filtering of Quantum Mechanical Processes.
Models and Control Policies for Biological Systems.

The Key-note Speakers are

Professor R.E. KALMAN, University of Florida, U.S.A., and Technische Hochschule, Zürich, Switzerland
Professor G. LEITMANN, University of California, Berkeley, U.S.A.
Professor S. MITTER, Massachusetts Institute of Technology, U.S.A.

Originally, it was thought that a gathering of a small number of specialists on an invited basis was sufficient for the purpose. However, the responses to our initial announcement surpassed our most optimistic estimate of the enthusiasm of workers in these areas. Subsequently, it was decided that we edit the Proceedings of this workshop as a book containing all the invited papers and selected contributed papers submitted to the workshop. This book will be published after the meeting by SPRINGER-VERLAG in the Series "Lecture Notes in Control and Information Sciences". The manuscripts contained in the present Preprints are extended summaries or full text of all papers available from authors at the time of the meeting.

This workshop could not have taken place without the technical and financial assistance of INRIA to whom we express our gratitude. In particular I would like to take this opportunity to thank his President Professor Alain BENSOUSSAN, the Director of the INRIA-SOPHIA ANTIPOLIS Research Center Professor Pierre BERNHARD, and the Director of Industrial and International Relations Georges NISSEN. I personnally address special thanks to Thérèse BRICHETEAU who, at the head of the Service des Relations Extérieures of INRIA, took care of all the myriad details of organization so efficiently and ably. We are most grateful to the expert assistance of Catherine JUNCKER who took care of the organization at Sophia and of the local arrangements. I am indebted to the Scientific Secretaries of the eighth INRIA Conference Analysis and Optimization of Systems for their help, in particular to Frédéric BONNANS for valuable advice in the preparation of the workshop. Especial thanks go to Gilbert MALLET who edited these Preprints. Last but not least, our thanks go to all the secretaries and technicists who handled the many problems at each step and at each level and contributed to the success of the third Bellman Continuum.

This international workshop was also financially supported by the Université Paris 7, one of the organizers of the meeting, and by the Association Française pour la Cybernétique Economique et Technique (AFCET), the Centre National de la Recherche Scientifique (CNRS), the European Research Office United States Army, the Ministère des Affaires Etrangères, the Ministère de l'Education Nationale, the Ministère de la Recherche et de l'Enseignement Supérieur, the United Nations Educational Scientific and Cultural Organization (UNESCO). Additional fellowships to the participants were provided by various organizations listed separately, to whom, as well as to the above mentioned organizations, we express our gratitude.

Austin BLAQUIERE (Chairman)

COMITE D'ORGANISATION

ORGANIZING COMMITTEE

A. BLAQUIERE,	Chairman, Université Paris 7, France
N. BELLMAN,	22 Latimer Road, Santa Monica, CA 90402, USA
A. BENSOUSSAN,	Université Paris-Dauphine / INRIA, France
P. BERNHARD,	INRIA-Sophia-Antipolis, France
Th. BRICHETEAU,	INRIA, France
A. ESOGBUE,	GA Institute of Technology, USA
G. FEICHTINGER,	Technical University Vienna, Austria
M. FLIESS,	Laboratoire des Signaux et Systèmes, CNRS-ESE, France
A. FOSSARD,	ENSAE, Toulouse, France
S. LEE,	Kansas State University, USA
G. LEITMANN,	University of California, Berkeley, USA
S. MEERKOV,	University of Michigan, USA
M. THOMA,	Technische Universität, Hannover, FRG
L. ZADEH,	University of California, Berkeley, USA

LISTE DES ORGANISMES APPORTANT LEURS CONCOURS FINANCIERS

LIST OF FINANCIAL SUPPORTERS

AFCET	Association Française pour la Cybernétique Economique et Technique
CNRS	Centre National de la Recherche Scientifique
ERO USA	European Research Office, United States Army
INRIA	Institut National de Recherche en Informatique et en Automatique
MAE	Ministère des Affaires Etrangères
MEN	Ministère de l'Education Nationale
MRES	Ministère de la Recherche et de l'Enseignement Supérieur
UNESCO	United Nations Educational, Scientific and Cultural Organization
UP7	Université Paris 7

Table des matières / Contents

COMMANDE DES SYSTEMES DYNAMIQUES INCERTAINS
CONTROL OF UNCERTAIN DYNAMICAL SYSTEMS

MODELISATION ET COMMANDE EN ECONOMIE
MODELS AND CONTROL POLICIES IN ECONOMICS

SYSTEMES STOCHASTIQUES ET QUANTIQUES

STOCHASTIC AND QUANTUM SYSTEMS

MODELISATION ET COMMANDE DES SYSTEMES BIOLOGIQUES ET DES ECOSYSTEMES

MODELS AND CONTROL POLICIES FOR BIOLOGICAL SYSTEMS AND ECOSYSTEMS

MATHEMATIQUES ET SYSTEMES, ASPECT CALCUL

MATHEMATICS AND SYSTEMS, COMPUTATIONAL BEARINGS

COMMANDE DES SYSTEMES DYNAMIQUES INCERTAINS

CONTROL OF UNCERTAIN DYNAMICAL SYSTEMS

CONTROLLING SINGULARLY PERTURBED UNCERTAIN DYNAMICAL SYSTEMS[1]

G. Leitmann
College of Engineering, University of California
Berkeley, California 94720 USA

INTRODUCTION

The prototype for the class of systems considered in this chapter is depicted in Figure 1 and consists of a dynamical process P (imperfectly known) controlled by a (judiciously designed) feedback law (operator F) acting on state data generated by sensor S and implemented via actuator A.

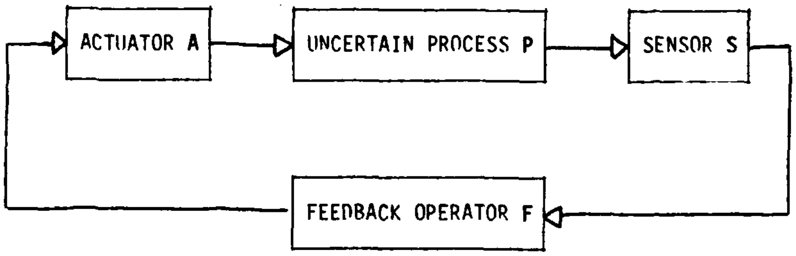


Figure 1. Prototype System

We assume (realistically) that the sensor and actuator are dynamic elements of the feedback loop; furthermore, we adopt the viewpoint that these dynamics are "fast" relative to those of the process P to be controlled. If this is not the case, then, at the modelling stage, the sensor and actuator should be explicitly incorporated as an integral part of the process to be controlled.

We recognize, of course, that in the context of nonlinear systems, the concept of "fas.ness" is difficult to quantify. Here we use the term loosely to indicate that the overall system exhibits a "two time scale" structure as described in the next section.

THE FULL-ORDER SYSTEM

The above prototype typifies a general class of singularly perturbed uncertain systems which can be decomposed, by means of a scalar parameter μ, into two coupled

[1] Based on research supported by the NSF and AFOSR. This paper deals with a special case of the problem considered in [8] and [9].

subsystems which henceforth will be referred to as the "slow" subsystem (with state $x(t)$) and the "fast" subsystem (with state $y(t)$). The parameter μ, henceforth referred to as the singular perturbation parameter, can be interpreted as some measure of the ratio of characteristic times of the fast and slow subsystems.

We model this general class of systems by the following coupled pair of differential equations.

$$\dot{x}(t) = X(t,x(t),y(\dot{t}),u(t)), \quad x(t)\in R^n, \quad u(t)\in R^m \tag{1a}$$

$$\mu\dot{y}(t) = Y(t,x(t),y(t),u(t),\mu), \quad y(t)\in R^p, \quad \mu\in(0,\infty) \tag{1b}$$

with measured output

$$z(t) = Sx(t) + Ty(t), \quad z(t)\in R^n \tag{1c}$$

where X and Y are uncertain functions with the following structure:

$$X(t,x,y,u) = A_{11}x + A_{12}y + B_1u + g_1(t,x,y,u) \tag{2a}$$

$$Y(t,x,y,u,\mu) = C(t)[A_{21}x + y + B_2u] + g_2(t,x,y,u,\mu) . \tag{2b}$$

A_{ij}, B_i, S and T are known constant real matrices; C is an uncertain measurable matrix-valued function; g_1 and g_2 are uncertain Caratheodory functions (i.e. measurable in their first argument, continuous in their other arguments and integrably bounded on compact sets).

Note that we require that the dimension of the output space coincides with the dimension of the slow subsystem state space. We refer to system (1)-(2) as the full-order system (a dynamical system on R^{n+p}).

Now suppose that the dynamics of the fast subsystem are neglected, i.e. suppose that μ is set to zero, in which case (1b) reduces to an algebraic constraint on (1a). This procedure yields the reduced-order system (a dynamical system on R^n). Suppose further that a feedback strategy is designed which guarantees some stability property P for the uncertain reduced-order system. (One such design is proposed in §5 and analysed in §6, using the deterministic framework developed in e.g. [1-7]). Then the essential question to be addressed is that of structural stability of property P with respect to singular perturbation, i.e. does property P persist when the fast dynamics are re-introduced? More usefully, does there exist a calculable threshold value $\mu^* > 0$ such that property P persists for all values of the singular perturbation parameter in the interval $(0,\mu^*)$?

Our objective is to answer such questions affirmatively, under additional hypotheses on the full-order system. The first of these is an assumption which ensures that a well-defined reduced order system results from setting $\mu = 0$ in (1b).

Assumption A1

(i) $C(\cdot) = C_0 + \Delta C(\cdot)$, where $C_0 \in R^{p\times p}$ is known with spectrum $\sigma(C_0) \subset \mathbb{C}^-$ (the open left half complex plane) and $\Delta C: R \to R^{p\times p}$ is an unknown measurable function with known bound κ_c (sufficiently small), viz. for all $t, \|\Delta C(t)\| \le \kappa_c < 1/2\|P\|^{-1}$, where $P > 0$ (symmetric) solves the Lyapunov equation $PC_0 + C_0^T P + I = 0$;

(ii) $g_2(\cdot,\cdot,\cdot,\cdot,0) = 0$.

THE REDUCED-ORDER SYSTEM

Solving the algebraic equation $Y(t,x,y,u,0) = 0$ for y (uniquely in view of Assumption A1) determines the function

$$(x,u) \mapsto H(x,u) \triangleq -[A_{21}x + B_2u] \ . \tag{3}$$

The reduced-order system associated with (1) is now defined as

$$\dot{x}(t) = X_r(t,x(t),u(t)), \qquad x(t) \in R^n \tag{4a}$$

with output

$$z(t) = Sx(t) + TH(x(t),u(t)), \qquad z(t) \in R^n \tag{4b}$$

where

$$X_r(t,x,u) \triangleq X(t,x,H(x,u),u) = \bar{A}x + \bar{B}u + \bar{g}(t,x,u) \tag{5a}$$

and

$$\bar{A} \triangleq A_{11} - A_{12}A_{21}, \quad \bar{B} \triangleq B_1 - A_{12}B_2, \quad \bar{g}(t,x,u) \triangleq g_1(t,x,H(x,u),u) \ . \tag{5b}$$

At this stage, we loosely define our preliminary goal as that of rendering, by feedback, some acceptably small compact neighborhood of the zero state of (4) globally attractive. Thus, it is not unreasonable to require the following of the nominal linear system pair $(\bar{A},\bar{B})$:

Assumption A2

(i) $(\bar{A},\bar{B})$ is a stabilizable pair,

(ii) $S - TA_{21}$ is non-singular.

Now, let $(Q,\gamma_0) \in R^{n\times n} \times R^+$ $(R^+ \triangleq [0,\infty))$ be a pair of design parameters with the properties (i) Q is symmetric and positive definite (ii) $\gamma_0 > 0$ if $\sigma(\bar{A}) \not\subset \mathbb{C}^-$.

These properties, in conjunction with A2, ensure that the Riccati equation

$$K\bar{A} + \bar{A}^T K + \mathcal{Q} - 2\gamma_0 K\bar{B}\bar{B}^T K = 0 \tag{6}$$

admits a unique real positive-definite symmetric solution $K > 0$. Hence, for example, in the absence of uncertainty ($\bar{g} \equiv 0$) and if $S = I$ and $T = 0$, the output feedback law $u = -\gamma_0 \bar{B}^T Kz$ renders the zero state of (4) asympstotically stable.

We now impose some additional structure and bounds on the system uncertainty.

Assumption A3

There exist known non-negative real numbers c_1, c_2, c_3, and unknown Caratheodory function $e: R \times R^n \times R^m \to R^m$ such that:

(i) $\bar{g} = \bar{B}e$;

and, for all $(t,x,u) \in R{\times}R^n{\times}R^m$,

(ii) $\|e(t,x,u)\| < c_1 + c_2 \|x\| + c_3 \|u\|$.

In the familiar terminology, the uncertainty is assumed to be matched and conebounded. The more general case of unmatched and non-conebounded uncertainty is considered in [8] and [9], albeit at the expense of a considerably more complicated controller design.

Define A: $P \to R^{n \times n}$ and $\Gamma_1, \Gamma_2 \subset R$ as follows:

$$A(\gamma) \overset{\Delta}{=} A_{21} - \gamma B_2 \bar{B}^T K \tag{7a}$$

$$\Gamma_1 \overset{\Delta}{=} \begin{cases} [\underline{\gamma}, \infty); & c_2 = 0 \\ (\underline{\gamma}, \infty); & c_2 > 0 \end{cases} ; \quad \underline{\gamma} \overset{\Delta}{=} (1 - c_3)^{-1}[\gamma_0 + c_2 \|\mathcal{Q}^{-1}\|] \tag{7b}$$

$$\Gamma_2 \overset{\Delta}{=} \{\gamma : |S - TA(\gamma)| \neq 0; \quad \kappa(\gamma) < (1 - 2\kappa_c\|P\|)/2\|PC_0\| + 2\kappa_c\|P\|)\} \tag{7c}$$

where

$$\kappa(\gamma) \overset{\Delta}{=} \gamma\|B_2 \bar{B}^T K[S - TA(\gamma)]^{-1} T\| . \tag{7d}$$

Then the following additional assumption is required.

Assumption A4

$$\Gamma^* \overset{\Delta}{=} \Gamma_1 \cap \Gamma_2 \neq \emptyset.$$

PROBLEM FORMULATION

Suppose a (time-dependent) output feedback control function $(t,z) \mapsto q(t,z)$ is designed which guarantees that the feedback-controlled reduced-order system (viz. $u(t) = -q(t,z(t))$ in (4)) possesses some desired stability property P, then the basic question to be addressed is that of robustness of P with respect to singular perturbation, where the singularly perturbed system is defined by (1) with $u(t) = -q(t,z(t))$; in particular, does there exist a (calculable) constant $\mu^* > 0$ such that the full system (1), under output feedback control $u(t) = -q(t,z(t))$, possesses property P for all values $\mu \in (0,\mu^*)$?

Here, we take the desired property P to be the existence of a compact set $\Sigma \subset R^n$ (respectively $\Sigma \subset R^{n+p}$) containing the origin which is a global uniform attractor for the reduced-order system (respectively, the full-order system) in the following sense.

Definition 1

A compact set $\Sigma \subset R^q$ is a global uniform attractor for the system

$$\dot{w}(t) = \Xi(t,w(t)), \quad w(t) \in R^q \qquad (*)$$

if the following properties hold:

(i) Existence and continuation of solutions: For each pair $(t_0,w^0) \in R \times R^q$ there exists a solution $w: [t_0,t_1) \to R^q$ (absolutely continuous function satisfying $(*)$ almost everywhere) with $w(t_0) = w^0$ and every such solution can be extended into a solution on $[t_0,\infty)$;

(ii) Uniform boundedness of solutions: For each $r > 0$ there exists $R(r) > 0$ such that $\|w(t)\| < R(r)$ for all t on every solution $w: [t_0,\infty) \to R^q$ of $(*)$ with $\|w(t_0)\| < r$, where $t_0 \in R$ is arbitrary;

(iii) Uniform stability of Σ: For each $d > 0$ there exists $D(d) > 0$ such that $w(t) \in \Sigma + dB$ for all t on every solution $w: [t_0,\infty) \to R^q$ of $(*)$ with $w(t_0) \in \Sigma + D(d)B$ where t_0 is arbitrary (note, B denotes the open unit ball in R^q and, for $\delta > 0$, $\Sigma + \delta B$ denotes the set $\{\sigma + \rho: \sigma \in \Sigma; \|\rho\| < \delta\}$);

(iv) <u>Global uniform attractivity of Σ</u>: For each $d > 0$ and $r > 0$ there exists $\tau(d,r) > 0$ such that $w(t) \in \Sigma + r\mathcal{B}$ for all $t > t_0 + \tau(d,r)$ on every solution $w: [t_0,\infty) \to R^q$ of $(^*)$ with $w(t_0) \in \Sigma + d\,\mathcal{B}$, where $t_0 \in R$ is arbitrary.

In the next section, we construct a feedback strategy which ensures property P for the reduced-order system (4).

NONLINEAR OUTPUT FEEDBACK

Choose ϵ_1, $\epsilon_2 > 0$; these are design parameters and can be chosen arbitrarily small. Define $p: R \times R^n \to R^m$ as

$$p(t,x) \triangleq p_0(x) + p_1(x) . \tag{8a}$$

The function p_0 is linear and is given by

$$p_0(x) \triangleq \gamma_1 \bar{B}^T Kx \tag{8b}$$

where $\gamma_1 \in R^+$ satisfies

$$\gamma_1 \in \Gamma^* . \tag{8c}$$

The function p_1 is nonlinear and bounded and is given by

$$p_1(x) \triangleq \begin{cases} \rho_1 \phi_1(\rho_1 \; \bar{B}^T Kx) & \text{if} \quad T = 0 \text{ or } B_2 = 0 \\ \\ 0 & \qquad\qquad \text{otherwise} \end{cases} \tag{8d}$$

where $\rho_1 \in R^+$ satisfies

$$\rho_1 \geq (1 - c_3)c_1 \tag{8e}$$

and $\phi_1: R^m \to R^m$ is any smooth (C^1) function which satisfies

$$\|\phi_1(v)\| \leq 1 , \quad \langle v, \phi_1(v)\rangle \geq \|v\| - \epsilon_1 \qquad \forall v \in R^m \tag{8f}$$

and which has bounded derivative $D\phi_1$; i.e., there exists $\kappa_\phi \in R^+$ such that $\|D\phi_1(v)\| \leq \kappa_\phi$ for all $v \in R^m$. The proposed output feedback control function $q: R \times R^n \to R^m$ is now defined by

$$q(t,z) \triangleq p(t, [S-TA(\gamma_1)]^{-1}z) . \tag{9}$$

Loosely speaking, the linear component (8b) of the control stabilizes (if necessary) the nominal linear system and counteracts part of the uncertainty e while nonlinear component (8d) (when active) counteracts the remaining part of e.

As an example of a function ϕ_1, satisfying the above requirements, consider the function

$$\phi_1 : v \mapsto \|v\| + \varepsilon_1]^{-1} v$$

for which (8f) clearly holds, and moreover, ϕ_1 is C^1 with $\|D\phi_1(v)\| < \varepsilon_1{}^{-1}$ for all $v \in R^m$.

A COMPACT ATTRACTOR FOR THE OUTPUT FEEDBACK CONTROLLED REDUCED-ORDER SYSTEM

For the reduced-order system (4), it may be verified that $q(t,z(t)) = p(t,x(t))$. Hence, setting $u(t) = -q(t,z(t))$ in (4a) yields the system

$$\dot{x}(t) = F_r(t,x(t)), \qquad x(t) \in R^n \tag{10a}$$

with

$$F_r(t,x) \triangleq \bar{A}x - \bar{B}p(t,x) + \bar{g}(t,x,-p(t,x)). \tag{10b}$$

As shown in [9], system (10) possesses stability property $\mathcal{P}$.

To this end, we define $V: R^n \to R^+$ (a Lyapunov function candidate) by

$$V(x) \triangleq \langle x, Kx \rangle \ . \tag{11}$$

Theorem 1.

There exists a closed ellipsoid

$$\Sigma_{r_0} \triangleq \{x \in R^n : \ V(x) < r_0^2\} \ ,$$

where r_0 is defined in [9], which is a global uniform attractor for system (10).

Our next objective is to show that property $\mathcal{P}$ is not destroyed by the reintroduction of the fast dynamics.

A COMPACT ATTRACTOR FOR THE OUTPUT FEEDBACK CONTROLLED FULL-ORDER SYSTEM

Define

$$h(x) \triangleq H(x, -p(t,x)) = -A(\gamma_1)x + B_2 p_1(x). \tag{12}$$

Our final assumption is now made.

Assumption A5

(i) For all (t,x),

$$\|g_1(t,x,y_1,-q(t,Sx+Ty_1)) - g_1(t,x,y_2,-q(t,Sx+Ty_2))\| < \lambda\|y_1-y_2\| \qquad \forall\, y_1,y_2$$

where $\lambda > 0$ is a known constant;

(ii) for all (t,x,y) and $\mu > 0$,

$$\|g_2(t,x,y,-q(t,Sx+Ty),\mu)\| < \mu[\kappa_1\|y-h(x)\| + \kappa_2\|x\| + \kappa_3]$$

where $\kappa_1, \kappa_2, \kappa_3 > 0$ are known constants.

While Assumptions 1 to 5 might appear somewhat esoteric, it is stressed that the class of systems which satisfy these hypotheses is far from trivial; for example, the assumptions hold for a class of uncertain systems with parasitic actuator and sensor dynamics considered in [10].

Let functions $F: R x R^n x R^p \to R^n$ and $G: R x R^n x R^p x R^+ \to R^p$ be given by

$$\begin{aligned} F(t,x,y) &\triangleq A_{11}x + A_{12}y - B_1q(t,Sx+Ty) + g_1(t,x,y,-q(t,Sx+Ty)) && (13)\\ &= F_r(t,x) + A_{12}[y-h(x)] + B_1[p(t,x)-q(t,Sx+Ty)]\\ &\quad + g_1(t,x,y,-q(t,Sx+Ty)) - g_1(t,x,h(x),-p(t,x)) \end{aligned}$$

$$\begin{aligned} G(t,x,y,\mu) &\triangleq C(t)[A_{21}x + y - B_2q(t,Sx+Ty)] + g_2(t,x,y,-q(t,Sx+Ty),\mu) && (14)\\ &= C(t)[y-h(x)] + C(t)B_2[p(t,x)-q(t,Sx+Ty)] + g_2(t,x,y,-q(t,Sx+Ty),\mu). \end{aligned}$$

Then the problem under consideration reduces to that of determining a threshold value $\mu^* > 0$ (if such exists) such that the system (two coupled subsystems):

$$\dot{x}(t) = F(t,x(t),y(t)) \qquad (15a)$$

$$\mu\dot{y}(t) = G(t,x(t),y(t),\mu) \qquad (15b)$$

possesses stability property P for all $\mu \in (0,\mu^*)$. We resolve this question via an analysis akin to that of [11].

As stated in [8] and shown in [9], the following theorem establishes property P for the full order system under output feedback control.

Theorem 2.

There exists a $\mu^* > 0$ such that, for all $\mu \in (0,\mu^*)$, a certain ellipsoid is a global uniform attractor for system (15); the value of μ^* and the definition of the attracting ellipsoid are given in [8] and [9]. Moreover, the reduced order dynamical behavior is recovered as $\mu \to 0$.[2]

EXAMPLE: UNCERTAIN SYSTEM WITH ACTUATOR AND SENSOR DYNAMICS

Consider the uncertain system

$$\dot{x}(t) = Ax(t) + [B + \Delta B(t)]y_1(t) + d(t,x(t)), \qquad x(t) \in R^n \tag{16a}$$

with actuator dynamics

$$\mu\dot{y}_1(t) = [C_1 + \Delta C_1(t)](y_1(t) - u(t)), \quad y_1(t),\ u(t) \in R^m \tag{16b}$$

and sensor dynamics

$$\mu\dot{y}_2(t) = [C_2 + \Delta C_2(t)](y_2(t) - x(t)), \quad y_2(t) \in R^n \tag{16c}$$

where the known nominal system matrices A, B, C_1, C_2 satisfy the following:

H1

(i) (A,B) is a stabilizable pair;

(ii) $\sigma(C_1) \subset \mathbb{C}^-$;

(iii) $\sigma(C_2) \subset \mathbb{C}^-$.

The uncertain functions $\Delta B(\cdot)$ and $d(\cdot,\cdot)$ are assumed to satisfy

H2

(i) $\Delta B(\cdot) = BE(\cdot)$, where $E(\cdot)$(unknown) is measurable with $\|E(t)\| \le \beta < 1\ \forall t$;

(ii) $d(\cdot,\cdot) = Bg(\cdot,\cdot)$, where $g(\cdot,\cdot)$ is a Caratheodory function with $\|g(t,x)\| \le \alpha_1\|x\| + \alpha_2 \quad \forall(t,x)$ and where α_1, α_2,β are known constants.

Let P (symmetric and positive definite) denote the unique solution of

$$P\begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix} + \begin{bmatrix} C_1^T & 0 \\ 0 & C_2^T \end{bmatrix} P + I = 0. \tag{17}$$

[2] Loosely speaking, in the sense that the projection of the attracting ellipsoid onto R^n approaches the attracting ellipsoid $\sum_{r_0}$ of the reduced order system.

Then the uncertain functions $\Delta C_1(\cdot)$ and $\Delta C_2(\cdot)$ are assumed to satisfy

H3

$\|\mathrm{diag}\{\Delta C_1(t),\ \Delta C_2(t)\}\| < \kappa_c < 1/2\|P\|^{-1}$ $\forall t$, where κ_c is a known constant.

The above can be interpreted in the context of system (1)-(2) by making the following identifications:

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \in R^p, \quad p \triangleq m + n \tag{18a}$$

$$A_{11} = A, \quad A_{12} = [B \vdots 0], \quad A_{21} = \begin{bmatrix} 0 \\ -I \end{bmatrix} \tag{18b}$$

$$B_1 = 0, \quad B_2 = \begin{bmatrix} -I \\ 0 \end{bmatrix}, \quad S = 0, \ T = [0 \vdots I] \tag{18c}$$

$$C(t) = C_0 + \Delta C(t), \quad C_0 = \mathrm{diag}\{C_1, C_2\}, \quad \Delta C(t) = \mathrm{diag}\{\Delta C_1(t), \Delta C_2(t)\} \tag{18d}$$

$$g_1(t,x,y,u) = d(t,x) + BE(t)[I \vdots 0]y \tag{18e}$$

$$g_2 \equiv 0 . \tag{18f}$$

In view of H1(ii),(iii) and H3, it is clear that Assumption A1 holds for this system.

Now,

$$\overline{A} = A_{11} - A_{12}A_{21} = A_{11} = A \tag{19a}$$

$$\overline{B} = B_1 - A_{12}B_2 = - A_{12}B_2 = B \tag{19b}$$

and hence, in view of H1(i), it follows that Assumption A2 holds.

Also,

$$H(x,u) = - [A_{21}x + B_2u] = \begin{bmatrix} u \\ x \end{bmatrix} \tag{20}$$

and

$$\overline{g}(t,x,u) = g_1(t,x,H(t,x),u) = \quad Be(t,x,u) \tag{21a}$$

where (21b)

$$e(t,x,u) = g(t,x) + E(t)u .$$

Thus, in view of H2, it is clear that Assumption A3 holds with $c_3 = \beta$. Proceeding,

$$A(\gamma) = A_{21} - \gamma B_2\bar{B}^T K = \begin{bmatrix} \gamma B^T K \\ -I \end{bmatrix} \tag{22a}$$

$$S - TA(\gamma) = I, \quad \kappa(\gamma) = \gamma\|B^T K\| \tag{22b}$$

$$\Gamma_1 = (-\infty, (1-2\kappa_c\|P\|)(2\|P C_0\| + 2\kappa_c\|P\|)^{-1}\|B^T K\|^{-1}) \subset R\ . \tag{22c}$$

Assumption A4 now reduces to the following:

$$A4^*: \ \underline{\gamma} < (1-2\kappa_c\|P\|)(1+2\kappa_c\|P\|)^{-1}\|B^T K\|^{-1} \quad .$$

Finally, it is readily verified that Assumption A5(ii) holds trivially (since $g_2 \equiv 0$) and A5(i) holds with $\lambda = \beta\|B\|$.

A specific example of this subclass of systems is considered in detail in [9].

OTHER METHODS

An approach, differing from the one proposed here, can be found in [12-15]. In these references, the design procedure requires the sequential construction of controllers which assure existence of global uniform attractors for (i) an approximation of the reduced order ("slow") subsystem, and (ii) the "fast" subsystem under the influence of the slow uncertainties. The controller for the full syst·m is then obtained as the sum of these subsystem controllers.

REFERENCES

[1] S. Gutman and G. Leitmann, "Stabilizing feedback control for dynamical systems with bounded uncertainty," Proc. IEEE Conference on Decision and Control (1976).

[2] G. Leitmann, "Deterministic control of uncertain systems," Astronautica Acta, 7(1980), pp. 1457-1461.

[3] G. Leitmann, "On the efficacy of nonlinear control in uncertain linear systems," J. Dynamic Systems Meas. Control, 103(1981), pp. 95-102.

[4] M. Corless and G. Leitmann, "Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems," IEEE Trans. Autom. Control, AC-26 (1981), pp. 1139-1144.

[5] B.R. Barmish and G. Leitmann, "On ultimate boundedness control of uncertain systems in the absence of matching conditions," IEEE Trans. Autom. Control, AC-27 (1982), pp. 153-158.

[6] B.R. Barmish, M. Corless and G. Leitmann, "A new class of stabilizing controllers for uncertain dynamical systems," SIAM J. Control and Optimization, 21 (1983), pp. 246-255.

[7] E.P. Ryan and M. Corless, "Ultimate boundedness and asymptotic stability of a class of uncertain dynamical systems via continuous and discontinuous feedback control," IMA J. Math. Control and Info., 1 (1984), pp. 223-243.

[8] G. Leitmann and E.P. Ryan, "Output feedback control of a class of singularly perturbed uncertain dynamical systems," Proceed. American Control Conference (1987), pp. 1590-1594.

[9] M. Corless, G. Leitmann and E.P. Ryan, "Control of uncertain systems with neglected dynamics," in "Variable Structure Control Systems", edited by A.I.S. Zinober, IEE Publ., London (in preparation).

[10] G. Leitmann, E.P. Ryan and A. Steinberg, "Feedback control of uncertain systems: robustness with respect to neglected actuator and sensor dynamics," Int. J. Control, 43 (1986), pp. 1243-1256.

[11] A. Saberi and H.K. Khalil, Quadratic-type Lyapunov functions for singularly perturbed systems," IEEE Trans. Autom. Contro, AC-29 (1984), pp. 542-550.

[12] F. Garofalo, "Composite control of a singularly perturbed uncertain system with slow uncertainties," Int. J. Control (to appear).

[13] F. Garofalo and G. Leitmann, "Nonlinear composite control of a nominally linear singularly perturbed uncertain system," Proceed. 12th IMACS World Congress (1988).

[14] F. Garofalo and G. Leitmann, "Nonlinear composite control of a class of nominally linear singularly perturbed uncertain systems," in "Variable Structure Control Systems", edited by A.I.S. Zinober, IEE Publ., London (in preparation).

[15] F. Garofalo and G. Leitmann, "Composite control of nonlinear, singularly perturbed uncertain systems," Proceed. Control 88, Oxford University (1988).

SOLUTIONS CONTINGENTES DE L'EQUATION D'HAMILTON-JACOBI-BELLMAN

Halina FRANKOWSKA
CEREMADE, Université de Paris-Dauphine
75775 Paris Cedex 16

Résumé

Il est connu que toute solution régulière d'une équation d'Hamilton-Jacobi-Bellman associée à un problème de contrôle optimal peut être utilisée pour la vérification de l'optimalité d'une trajectoire du système, ainsi que pour la construction des rétroactions optimales.

En général de telles solutions régulières n'existent pas et on introduit les solutions généralisées (solutions de viscosité ou autres).

Dans cet exposé on pose la question suivante : quelles sont des conditions nécessaires et suffisantes pour qu'une fonction $V : R \times R^n \longrightarrow R \cup \{\pm\infty\}$ vérifie les propriétés suivantes :

1) V est monotone le long des trajectoires du système
2) V est constante le long d'au moins une trajectoire (qui est une solution optimale du problème).

Les propriétés 1) et 2) sont cruciales pour l'application des techniques de vérification.

On démontre aussi que de telles fonctions forment une sous-classe des solutions de viscosité de l'équation d'Hamilton-Jacobi-Bellman.

Les propriétés 1) et 2) sont toujours vérifiées par la fonction valeur associée au problème. Mais cette dernière étant souvent discontinue, il est intéressant de trouver d'autres fonctions qui satisfassent 1) et 2).

La deuxième partie de l'exposé concerne la construction des rétroactions optimales associées à de telles fonctions.

On Robust Control of Uncertain Linear Systems in the Absence of Matching Conditions

Harold Stalford

Aerospace and Ocean Engineering
Interdisciplinary Center for Applied Mathematics
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT

We establish a general robust control result for linear time-invariant uncertain systems using the Lyapunov approach initiated by Leitmann and Gutman. We show that systems satisfying matching conditions are handled by this result. We give necessary and sufficient conditions for the existence of a robust sliding mode controller. We show that its existence implies the existence of a robust linear controller. A counter example is provided to establish that the converse does not hold. The feedback controllers treated are functions of the complete state without any dynamic compensation.

1. INTRODUCTION

The Lyapunov approach to uncertain systems received an initial thrust by Leitmann and Gutman, [1] - [7], for systems satisfying matching conditions. They are joined by numerous authors (e.g. [8] - [33]) in extending the Lyapunov approach to handle more general systems since it is well suited for addressing structured uncertainty. Our work herein focuses on applying the Lyapunov approach to systems which have constant uncertainties but do not necessarily satisfy the matching conditions. It builds on the work of [9], [14], and [20] - [33]. Our main objective is to establish a robust control result based on the Lyapunov approach which generalizes some of the past work on linear uncertain systems with constant uncertainties. We specifically consider linear and sliding mode controllers and give necessary and sufficient conditions for their existence. We prove that the existence of a robust stabilizing sliding mode controller implies the existence of a robust stabilizing linear controller. The converse does not hold. We provide a counter example showing the existence of a robust linear controller in the absence of such a sliding mode controller. Herein, we use the term stability to mean that the poles are in the left-half plane, i.e., asymptotic stability or, equivalently, that the characteristic polynomial is Hurwitz. We say that a controller is robust if it asymptotically stabilizes the system for all uncertainties. We treat both the scalar input and the multi-input problems.

We investigate the robust control of linear time-invariant uncertain systems that are not required necessarily to satisfied matching conditions:

$$\dot{x} = A(\gamma)x + B(\gamma)u, \quad \gamma \in \Gamma \tag{1}$$

where $A(\gamma)$ is a nxn uncertain matrix, $B(\gamma)$ is an nxm uncertain matrix with full rank ($m \leq n$) and γ belongs to a set of uncertainties Γ where Γ is a simply connected, compact subset of p-dimensional Euclidean space E^p. We assume that $A(\gamma)$ and $B(\gamma)$ are continuous with respect to the uncertainty argument $\gamma \in \Gamma$. In this paper we consider only full state feedback controllers $u(x)$, i.e., those which are functions of the state x only. That is, we do not address dynamic compensation as part of the feedback controller. We require that system (1) satisfy the controllability assumption:

ASSUMPTION I. For each $\gamma \in \Gamma$ the pair $(A(\gamma), B(\gamma))$ is controllable.

The controllability assumption is equivalent to the assumption that closed-loop poles can be arbitrarily placed by a suitable gain matrix. We state this equivalent assumption:

ASSUMPTION I'. For each $\gamma \in \Gamma$ and prescribed eigenvalues $\Lambda(\gamma) = (\lambda_1(\gamma), \ldots, \lambda_n(\gamma))$ in which imaginary eigenvalues occur in complex conjugate pairs there exists a real gain matrix $K(\gamma)$ such that the closed-loop matrix

$$\bar{A}(\gamma) = A(\gamma) - B(\gamma)K(\gamma) \tag{2}$$

has the prescribed eigenvalues $\Lambda(\gamma)$.

For arbitrarily prescribed eigenvalues $\Lambda(\gamma)$, $\gamma \in \Gamma$, we can rewrite (1) as

$$\dot{x} = \bar{A}(\gamma)x + B(\gamma)[K(\gamma)x + u] \tag{3}$$

where $K(\gamma)$ is the corresponding gain matrix and $\bar{A}(\gamma)$ satisfies (2).

The next assumption makes it possible to define a control law with which to stabilize (1) in the presence of uncertainties $\gamma \varepsilon \Gamma$.

ASSUMPTION II. For each $\gamma \varepsilon \Gamma$ there exist an mxn gain matrix $K(\gamma)$. an invertible mxm matrix $R(\gamma)$ and an nxn symmetric, positive definite matrix $Q(\gamma)$ such that

(*i*) $\overline{A}(\gamma) = A(\gamma) - B(\gamma)K(\gamma)$ is asymptoticaliy stable

(*ii*) $F = R^{-1}(\gamma)B^T(\gamma)P(\gamma)$ is a constant mxn matrix where $P(\gamma)$ is the symmetric, positive definite solution of Lyapunov equation

$$P(\gamma)\overline{A}(\gamma) + \overline{A}^T(\gamma)P(\gamma) + Q(\gamma) = 0 \tag{4}$$

We make the following assumption on the mxm matrix $R(\gamma)$ which is defined in Assumption II.

ASSUMPTION III. For $\gamma \varepsilon \Gamma$ the matrix $\Phi(\gamma)$ defined as

$$\Phi(\gamma) = \frac{R^T(\gamma) + R(v)}{2} \tag{5a}$$

is positive definite and has the square root form

$$\Phi(\gamma) = S^T(\gamma)S(\gamma) \tag{5b}$$

where $S(\gamma)$ is invertible. The following upper bound exists and is finite

$$h = \max_{\gamma \varepsilon \Gamma}\left[\|S^{-1}(\gamma)\| \ \|S^{-T}(\gamma) \ R^T(\gamma)\|\right] \tag{6}$$

In Sections 2-4 and 6 we show how to use the constant matrix F in establishing a robust controller.

2. MAIN ROBUST CONTROL RESULT

Assumptions I - III permit the development of a robust control law that is discontinuous in nature. This is established in the next theorem.

THEOREM 1: If system (1) satisfies Assumptions I - III then the discontinuous controller

$$u(x) = -\frac{Fx}{\|Fx\|}\,\rho(x), \quad Fx \neq 0 \tag{7}$$

stabilizes (1) for all $\gamma \varepsilon \Gamma$ where $\rho(x)$ satisfies

$$\rho(x) = h \max_{\gamma \varepsilon \Gamma} \|K(\gamma)\,x\| \tag{8}$$

The scalar h is given by (6) and the gain matrix $K(\gamma)$ is defined in Assumption II.

PROOF: For $\gamma \varepsilon \Gamma$ let $K(\gamma)$, $R(\gamma)$, $Q(\gamma)$, $P(\gamma)$ and F be the matrices described in Assumption II. Define the Lyapunov function

$$V(\gamma) = x^T P(\gamma)x \tag{9}$$

It has the time derivative

$$\dot{V}(\gamma) = -x^T Q(\gamma)x + 2[B^T(\gamma)P(\gamma)x]^T\,[K(\gamma)x + u] \tag{10}$$

Using property *(ii)* of Assumption II this derivative becomes

$$\dot{V}(\gamma) = -x^T Q(\gamma)x + 2[Fx]^T\,R^T(\gamma)\,[K(\gamma)x + u(x)] \tag{11}$$

We show that the control law (7) yields

$$\dot{V}(\gamma) \leq -x^T Q(\gamma)x, \quad \gamma \varepsilon \Gamma \tag{12}$$

Since $Q(\gamma) > 0$ (i.e., positive definite) it suffices to show that $W(\gamma)$ is nonpositive:

$$W(\gamma) \stackrel{\cdot}{=} 2[Fx]^T R^T(\gamma)[K(\gamma)x + u(x)] \leq 0 \tag{13}$$

Consider a control law of the form (7) in which the scalar function $\rho(x)$ is defined by (8). substitution of (7) into (13) yields

$$W(\gamma) = 2W_1(\gamma) - 2W_2(\gamma) \leq 0 \tag{14}$$

where

$$W_1(\gamma) = [Fx]^T R^T(\gamma)\, K(\gamma) x \qquad (15a)$$

$$W_2(\gamma) = [Fx]^T R^T(\gamma) \frac{Fx}{\|Fx\|} \rho(x), \quad Fx \neq 0 \qquad (15b)$$

Eq. (15b) can be rewritten as

$$W_2(\gamma) = [Fx]^T \Phi(\gamma) \frac{Fx}{\|Fx\|} \rho(x) \qquad (16)$$

or, equivalently as,

$$W_2(\gamma) = [S(\gamma) Fx]^T \frac{S(\gamma)\, Fx}{\|Fx\|} \rho(x) \qquad (17)$$

where $\Phi(\gamma)$ and $S(\gamma)$ are defined by (5) and (6). Making the vector definition

$$y(\gamma) = S(\gamma)\, Fx \qquad (18)$$

Eq. (17) becomes

$$W_2(\gamma) = \frac{y^T(\gamma)\, y(\gamma)}{\|Fx\|} \rho(x) \qquad (19)$$

Eq. (15a) can be rewritten as

$$W_1(\gamma) = y^T(\gamma) z(\gamma) \qquad (20)$$

where

$$z(\gamma) = S(\gamma)[\Phi(\gamma)]^{-1} R^T(\gamma)\, K(\gamma) x \qquad (21)$$

Inequality (14) is met provided

$$W_1(\gamma) \leq W_2(\gamma) \qquad (22)$$

In terms of (19) and (20) this inequality is given by

$$y^T(\gamma) z(\gamma) \leq \frac{y^T(\gamma) y(\gamma)}{\|Fx\|} \rho(x), \quad Fx \neq 0 \qquad (23)$$

This inequality is met provided

$$\|z(\gamma)\| \leq \frac{\|y(\gamma)\|}{\|Fx\|} \rho(x), \quad Fx \neq 0 \qquad (24)$$

Taking the norm of (21) yields

$$\|z(\gamma)\| \leq \|S^{-T}(\gamma)\, R^T(\gamma)\| \, \|K(\gamma) x\| \qquad (25)$$

Multiplying both sides by the norm $\|S^{-1}(\gamma)\|$ gives

$$\|S^{-1}(\gamma)\| \, \|z(\gamma)\| \leq \rho(x) \qquad (26)$$

Observe that

$$\|Fx\| = \|S^{-1}(\gamma) S(\gamma) Fx\| \leq \|S^{-1}(\gamma)\| \, \|y(\gamma)\| \qquad (27)$$

from which it follows that

$$1 \le \frac{\|S^{-1}(\gamma)\| \, \|y(\gamma)\|}{\|Fx\|}, \quad Fx \ne 0 \tag{28}$$

Multiplying both sides by $\rho(x)$ yields

$$\rho(x) \le \frac{\|S^{-1}(\gamma)\| \, \|y(\gamma)\|}{\|Fx\|} \rho(x), \quad Fx \ne 0 \tag{29}$$

The inequalities (26) and (29) yield

$$\|z(\gamma)\| \le \frac{\|y(\gamma)\|}{\|Fx\|} \rho(x), \quad Fx \ne 0 \tag{30}$$

This verifies (24) which establishes (12). By the theory of Lyapunov, the control law (7) stabilizes (1) for each uncertainty $\gamma \in \Gamma$.

3. ROBUST CONTROL IN THE PRESENCE OF MATCHING CONDITIONS

Systems which satisfy the matching conditions of linear uncertain systems, [2] - [7], satisfy Assumptions I - III. This result is given by the next theorem.

THEOREM 2: Let system (1) satisfy the following matching conditions: There exist an nxn matrix A and an nxm matrix B and for each $\gamma \in \Gamma$ there exist an mxn gain matrix $D(\gamma)$ and an invertible mxm matrix $\Pi(\gamma)$ such that

(a) $A(\gamma) = A + BD(\gamma)$.

(b) $B(\gamma) = B\Pi(\gamma)$.

(c) (A, B) is a controllable pair

(d) $\Phi(\gamma)$ is an mxm positive definite matrix where

$$\Phi(\gamma) = \frac{\Pi^T(\gamma) + \Pi(\gamma)}{2} \tag{31}$$

Then Assumptions I - III are met. As a consequence of Theorem 1, there exists a robust stabilizing control law of the form

$$u = -Kx + \frac{Fx}{\|Fx\|} \rho(x) \tag{32}$$

such that

$$\bar{A} = A - BK \tag{33}$$

is asymptotically stable and such that

$$F = B^T P \tag{34}$$

where P is the symmetric, positive definite solution of the Lyapunov equation

$$P\bar{A} + \bar{A}^T P + Q = 0 \tag{35}$$

in which $Q > 0$ is arbitrarily chosen.

PROOF: Conditions (a) - (c) imply that $(A(\gamma), B(\gamma))$ is controllable for $\gamma \in \Gamma$. Controllability is invariant under linear feedback and coordinate transformation on the input, [34]. Thus Assumption I is met. Since (A, B) is controllable there exists a gain matrix K such that $\bar{A}$ of (33) is asymptotically stable. Define the uncertain gain matrix

$$K(\gamma) = \Pi^{-1}(\gamma)[D(\gamma) + K] \tag{36}$$

Using conditions (a) and (b) we find $\bar{A}(\gamma)$ of condition (*i*) of Assumption II reduces to

$$\bar{A}(\gamma) = A - BK \tag{37}$$

and is, therefore, asymptotically stable for $\gamma \in \Gamma$. Select any $Q > 0$. Let P be the solution of (36) and let F be defined by (34). For $\gamma \in \Gamma$ define

$$R(\gamma) = \Pi^T(\gamma) \tag{38}$$

The matrix F of condition (*ii*) of Assumption II and that of (34) are identical. That is, (34) can be rewritten as

$$F = \Pi^{-T}(\gamma)[B\,\Pi(\gamma)]^T P \tag{39}$$

which, in view of condition (b) and (38), is equivalent to

$$F = R^{-1}(\gamma)\,B^T(\gamma)\,P \tag{40}$$

Thus, condition (*ii*) of Assumption II is met with

$$P(\gamma) = P \tag{41}$$

Condition (d) implies Assumption III since $B(\gamma)$ is continuous and Γ is compact. That is, h exists and is finite. Since all conditions of Theorem 1 are met, the existence of the stabilizing control law (32) follows with

$$\rho(x) = h \max_{\gamma \in \Gamma} \|K(\gamma)x\| \tag{42}$$

where $K(\gamma)$ is defined by (36).

4. ROBUST CONTROL IN THE ABSENCE OF MATCHING CONDITIONS: SCALAR INPUT

We show that the robust control assumptions presented in [29] for scalar control satisfy the assumptions of Theorem 1.

Consider system (1) with scalar control. The input matrix $B(\gamma)$ is a column vector. The work in [29] assumes that the system (1) is controllable. Assumption I. Under this assumption there is a unique coordinate transformation $T(\gamma)$

$$z = T(\gamma)x \tag{43}$$

of (1) to the following controllable companion form, [34],

$$\dot{z} = A_z(a(\gamma))z + B_z u(x) \tag{44}$$

where

$$A_z(a(\gamma)) = \begin{bmatrix} -a_1(\gamma) & -a_2(\gamma) & \dots & -a_{n-1}(\gamma) & -a_n(\gamma) \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \tag{45a}$$

$$A_z(a(\gamma)) = T(\gamma)\,A(\gamma)\,T^{-1}(\gamma) \tag{45b}$$

and

$$B_z = [1, 0, 0, \dots, 0]^T \tag{46a}$$

$$B_z = T(\gamma)\,B(\gamma) \tag{46b}$$

The vector $\mathrm{a}(\gamma) = (a_1(\gamma), \dots, a_n(\gamma))$ is the coefficient vector of the open-loop characteristic polynomial:

$$a_\gamma(s) = \det[sI - A(\gamma)] \tag{47}$$

We need the following definition in order to introduce the next assumption of [29]-[31].

DEFINITION 1: The row vector $P_1 = (P_{11}, P_{12}, \dots, P_{1n})$ is said to be $n-1$ stable provide $P_{11} > 0$ and the polynomial

$$P_{11}\,\lambda^{n-1} + P_{12}\,\lambda^{n-2} + \dots + P_{1n} = 0 \tag{48}$$

is Hurwitz (i.e., all eigenvalues are in left-half plane).

ASSUMPTION IV: There exist an uncetain $\gamma_0 \,\varepsilon\, \Gamma$ and an $n-1$ stable row vector $P_1(\gamma_0)$ such that

$$P_1(\gamma) = P_1(\gamma_0)\,T(\gamma_0)\,T^{-1}(\gamma) \tag{49}$$

is $n-1$ stable for all $\gamma \,\varepsilon\, \Gamma$

The concept of a vector being $n-1$ stable is fundamental in the asymptotically stable solution of Lyapunov equation. This result is presented in the next lemma. Its proof is given in [30].

LEMMA 1. Let $a = (a_1, \dots, a_n)$. Define $A(a)$ to be in the controllable companion form (45). Let P be the solution to the Lyapunov equation

$$P\,A(a) + A^T(a)\,P + Q = 0 \tag{50}$$

where $Q > 0$ and $Q = Q^T$. Then $A(a)$ is stable if, and only if, P_1 is $n-1$ stable where P_1 is the first row of P.

PROOF: See [30].

The next lemma is a consequence of Lemma 1.

LEMMA 2. Suppose Assumption IV holds. For each $\gamma \epsilon \Gamma$ define $Q(\gamma) > 0, Q(\gamma) = Q^T(\gamma)$. Then for each γ, there is a unique stable coefficient vector $\hat{a}(\gamma)$ satisfying Lyapunov equation

$$P(\gamma)A(\hat{a}(\gamma)) + A^T(\hat{a}(\gamma))P(\gamma) + Q(\gamma) = 0 \tag{51}$$

where $P_1(\gamma)$, the first row of $P(\gamma)$, is prescribed under Assumption IV. That is, $A(\hat{a}(\gamma))$ is stable for $\gamma \epsilon \Gamma$.

The above lemmas are used in the next theorem to establish a stabilizing controller for system (1).

THEOREM 3. If Assumptions I and IV hold then there is a stabilizing controller for system (1) having the form

$$u = -\frac{Fx}{\|Fx\|}\rho(x), \quad Fx \neq 0 \tag{52}$$

where F is a constant row vector and $\rho(x)$ is a nonnegative scalar function of the state x.

PROOF: Since system (1) is controllable for each uncertainty $\gamma\epsilon\Gamma$ it can be transformed to the controllable companion form (44). Assumption IV implies there is a stable coefficient vector $\hat{a}(\gamma)$ for $\gamma \epsilon \Gamma$ such that (51) is satisfied. Define $\sigma(\gamma)$ to be the difference between the stable coefficient vector $\hat{a}(\gamma)$ and the open-loop characteristic polymonial coefficient vector $a(\gamma)$ of System (1)

$$\sigma(\gamma) = \hat{a}(\gamma) - a(\gamma) \tag{53}$$

Note that the negative of $a(\gamma)$ is contained in the first row of (45). Substitution of (53) into (44) yields

$$\dot z = A_z(\hat a(\gamma))z + B_z[\sigma(\gamma)T(\gamma)x + u(x)] \tag{54}$$

after making use of (43). We use the symmetric, positive definite solution $P(\gamma)$ of (51) to construct the Lyapunov function

$$V(\gamma) = z^T P(\gamma) z \tag{55}$$

Taking its derivative gives

$$\dot V(\gamma) = -z^T Q(\gamma)z + 2[Fx]^T[\sigma(\gamma)\,T(\gamma)x + u(x)] \tag{56}$$

where F satisfies

$$F = P_1(\gamma_0)T(\gamma_0) \tag{57a}$$

and as a consequence of Assumption IV we have

$$F = P_1(\gamma)T(\gamma) \tag{57b}$$

or, equivalently,

$$F = B_z^T P(\gamma)T(\gamma) \tag{57c}$$

where $P(\gamma)$ satisfies (51) and $T(\gamma)$ satisfies (43). Any admissible control law $u(x)$ satisfying

$$u(x) \le -\max_{\gamma \epsilon \Gamma}[\sigma(\gamma)T(\gamma)x]\ , \quad Fx > 0 \tag{58a}$$

$$u(x) \ge \max_{\gamma \epsilon \Gamma}[\sigma(\gamma)T(\gamma)x]\ , \quad Fx < 0 \tag{58b}$$

stabilizes (1) since for such a control law

$$\dot V(\gamma) \le -z^T Q(\gamma)z\,, \quad \gamma \ \epsilon\ \Gamma \tag{59}$$

The maxima of (58) exist since Γ is compact and since the functions $\sigma(\gamma)$ and $T(\gamma)$ are continuous on Γ. An admissible control law satisfying (58) is (52) where

$$\rho(x) = \max_{\gamma \epsilon \Gamma} \|\sigma(\gamma)\,T(\gamma)\,x\| \tag{60}$$

and F is given by (57). In the next theorem we establish that a system satisfying Assumption IV also satisfies Assumption II.

THEOREM 4: If the system (1) satisfies Assumptions I and IV then Assumptions II and III are met.

PROOF: We make the following identifications

$$\overline{A}(\gamma) = T^{-1}(\gamma)\,A(\hat{a}(\gamma))\,T(\gamma) \tag{61a}$$

$$\overline{P}(\gamma) = T^{T}(\gamma)\,P(\gamma)\,T(\gamma) \tag{61b}$$

$$\overline{Q}(\gamma) = T^{T}(\gamma)\,Q(\gamma)\,T(\gamma) \tag{61c}$$

$$K(\gamma) = \sigma(\gamma)\,T(\gamma) \tag{61d}$$

where $T(\gamma)$ is defined by (43), where $P(\gamma)$, $Q(\gamma)$and$A(\hat{a}(\gamma))$ are defined by (51) and where $\sigma(\gamma)$ is defined by (53). The matrix $A(\hat{a}(\gamma))$ is asymptotically stable. This follows from Lemma 2 and the fact that eigenvalues are invariant under coordinate transformation. From (44), (45), (53) and (54) it follows that

$$\overline{A}(\gamma) = A(\gamma) - B(\gamma)\,K(\gamma) \tag{62}$$

so that condition (i) of Assumption III is met. The vector F of (57) satisfies

$$F = B^{T}(\gamma)\,\overline{P}(\gamma) \tag{63}$$

where $\overline{P}(\gamma)$ is the solution of the Lyapunov equation

$$\overline{P}(\gamma)\,\overline{A}(\gamma) + \overline{A}^{T}(\gamma)\,\overline{P}(\gamma) + \overline{Q}(\gamma) = 0 \tag{64}$$

which shows that condition (ii) of Assumption II is met. Here, the scalar R = 1. Thus Assumption III is also met.

Theorems 3 and 4 establish that Assumption IV implies Assumption II. The converse need not hold. Thus Assumption IV is a stronger assumption. Assumption IV admits a sliding mode controller (52). From the next theorem we [illegible] that it also admits a stabilizing linear controller.

THEOREM 5: If Assumption I and IV hold then there exists a stabilizing linear control

$$u = -cFx \tag{65}$$

where F is defined as in Theorem 3 and the scalar c satisfies

$$c > \frac{1}{2}\max_{\gamma\in\Gamma}\|Q^{-1}(\gamma)\|\;\max_{\gamma\in\Gamma}\|K(\gamma)\|^{2} \tag{66}$$

where $Q(\gamma)$, $\gamma \in \Gamma$, is defined as in Lemma 2 and where $K(\gamma)$ is given by (61d).

PROOF: See [31].

The maxima of (66) exist since $K(\gamma)$ is continuous, Γ is compact and the matrices $Q(\gamma)$ are chosen in a continuous manner. Usually $Q(\gamma)$ is set to be the identity I or it is computed from

$$Q(\gamma) = T^{-T}(\gamma)\ Q\, T^{-1}(\gamma) \tag{67}$$

where Q is a prescribed symmetric, positive definite matrix. The next result gives an equivalence between Assumption IV and a minimum phase condition on the system.

THEOREM 6: Assumption IV is met if, and only if, there is a row vector F such that

$$F[sI - A(\gamma)]^{-1}\ B(\gamma), \quad \gamma \in \Gamma \tag{68}$$

is minimum phase with n-1 transmission zeros where I is the nxn identity matrix. That is, the determinant

$$\det\begin{bmatrix} A(\gamma) - sI & B(\gamma) \\ F & 0 \end{bmatrix} = 0, \quad \gamma \in \Gamma \tag{69}$$

is Hurwitz.

PROOF: From (49) of Assumption IV

$$P_1(\gamma) = F\,T^{-1}(\gamma), \quad \gamma \in \Gamma \tag{70}$$

where $P_1(\gamma)$ is n-1 stable with polynomial Eq. (48) can be rewritten as

$$P_1(\gamma)[s^{n-1}\ s^{n-2} \ldots s\ \ 1]^T = 0 \tag{71}$$

where $s = \sigma + j\omega = \lambda$. Multiplying (70) on both sides by $[s^{n-1}\ s^{n-2} \ldots s\ \ 1]^T$ gives

$$FT^{-1}(\gamma)[s^{n-1}\ s^{n-2} \ldots s\ \ 1]^T = 0 \tag{72}$$

The open-loop characteristic polynomial $a_\gamma(s)$, (47), is given by

$$a_\gamma(s) = s^n + a_1(\gamma)s^{n-1} + \ldots + a_{n-1}(\gamma)s + a_n(\gamma) = 0 \tag{73}$$

Since $A_s(a(\gamma))$andB_s are in the controller companion form (45) and (46) we have the following identity from linear system theory, [34]:

$$[sI - A_s(a(\gamma))]^{-1}B_s = \frac{[s^{n-1}s^{n-2} \ldots s\ 1]^T}{a_\gamma(s)} \tag{74}$$

Substitution from (45b) and (46b) into (74) gives

$$T(\gamma)[sI - A(\gamma)]^{-1}B(\gamma) = \frac{[s^{n-1}s^{n-2} \dots s\ 1]^T}{a_\gamma(s)} \tag{75}$$

Multiplying both sides by $FT^{-1}(\gamma)$ yields

$$F[sI - A(\gamma)]^{-1}B(\gamma) = \frac{FT^{-1}(\gamma)[s^{n-1}s^{n-2} \dots s\ 1]^T}{a_\gamma(s)} = 0 \tag{76}$$

after making use of (72). The transmission zeros, [35] , of (76) are the n-1 stable eigenvalues of the (n-1) stable $P_1(\gamma)$ row vector of (70). This proves that (68) is minimum phase. From (76) we have

$$\det[sI - A(\gamma)]\ F[sI - A(\gamma)]^{-1}B(\gamma) = 0 \tag{77}$$

A reciprocal form of (77) is given by, [34],

$$\det\begin{bmatrix} sI - A(\gamma) & B(\gamma) \\ -F & 0 \end{bmatrix} = 0 \tag{78}$$

which yields (69). Since $P_1(\gamma)$ is n-1 stable it follows that (71) is Hurwitz. Thus (69) is Hurwitz.

Conversely, if there exists an F such that (69) is Hurwitz then the vector $P_1(\gamma)$ defined by (70) is n-1 stable and Assumption IV is met. From the above theorem we have the corollary.

COROLLARY 1. A necessary and sufficient conditions for the existence of a stabilizing sliding mode controller

$$u = -\frac{Fx}{\|Fx\|}\rho(x), \quad Fx \neq 0 \tag{79}$$

of (1) is the existence of a row-vector F such that (69) is Hurwitz for all $\gamma \varepsilon \Gamma$.

The existence of a stabilizing linear controller

$$u = -Kx \tag{80}$$

does not imply the existence of a stabilizing sliding mode controller (79). Before this is illustrated by an example we give necessary and sufficient conditions for the existence of a linear controller (80).

THEOREM 7: A necessary and sufficient condition that there exist a stabilizing linear controller (80) of system (1) is that there exists a row vector K such that the following determinant is Hurwitz:

$$\det\begin{bmatrix} sI - A(\gamma) & B(\gamma) \\ -K & 1 \end{bmatrix} = 0, \quad \gamma \ \varepsilon \ \Gamma \tag{81}$$

PROOF: Suppose there is a row vector K such that (80) asymptotically stabilizes (1). The feedback matrix

$$A_c(\gamma) = A(\gamma) - B(\gamma)K, \quad \gamma \ \varepsilon \ \Gamma \tag{82}$$

is asymptotically stable and the determinant

$$a_\gamma^c(s) = \det[sI - A_c(\gamma)] = 0 \tag{83}$$

is Hurwitz. Eq. (83) can be written as the following series of identities, [34],

$$a_\gamma^c(s) = \det\{[sI - A(\gamma)]\ [I + [sI - A(\gamma)]^{-1}B(\gamma)K]\} \tag{84a}$$

$$a_\gamma^c(s) = a_\gamma(s)\ \det[I + [sI - A(\gamma)]^{-1}B(\gamma)K] \tag{84b}$$

$$a_\gamma^c(s) = a_\gamma(s)\ [1 + K[sI - A(\gamma)]^{-1}B(\gamma)] \tag{84c}$$

where $a_\gamma(s)$ is the open-loop characteristic polynomial (47). The reciprocal form of (81) is (84c), [34]. That is, (81) and (84c) are identities. Therefore, (81) is Hurwitz if, and only if, (83) is Hurwitz. Eq. (84c) can be used to prove Theorem 5. If (76) is Hurwitz then with

$$K = cF \tag{85}$$

Eq. (84c) becomes

$$a_\gamma^c(s) = a_\gamma(s)[1 + c\,F[sI - A(\gamma)]^{-1}B(\gamma)] \tag{86}$$

which is Hurwitz for sufficiently large c. That is, in view of (71) -(76), Eq. (86) can be rewritten as

$$a_\gamma^c(s) = s^n + c\,P_1(\gamma)[s^{n-1}, s^{n-2} \ldots s\ \ 1]^T + [a_\gamma(s) - s^n] \tag{87}$$

in which the last term is an n-1 order polynomial that is dominated by the middle term for large c. The first two terms give a Hurwitz polynomial for sufficiently large c. As a consequence, the existence of a robust stabilizing sliding mode controller (52) implies the existence of a robust stabilizing linear controller (65). In general, the converse does not hold as is illustrated by the following example.

5. EXAMPLE OF ROBUST LINEAR CONTROLLER WITHOUT SLIDING MODE CONTROLLER

Consider the uncertain system

$$\dot{x} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} x + \begin{bmatrix} \gamma \\ 1 \end{bmatrix} u, \quad \gamma \varepsilon \Gamma \tag{88}$$

where $\Gamma = [-M, M]$ and where M is a positive scalar greater than 1

$$M \geq 1 \tag{89}$$

The determinant of the controllability matrix $[B(\gamma), AB(\gamma)]$ is given by $\gamma^2 + 1$ which satisfies the equality

$$\gamma^2 + 1 > 0 \quad \forall \gamma \varepsilon (-\infty, \infty) \tag{90}$$

The system (88) is controllable for all uncertainties γ. Thus Assumption I is satisfied. The requirement for the existence of a stable sliding mode surface

$$Fx = 0 \tag{91}$$

depends on (69) being Hurwitz. For our example system (88) Eq. (69) reduces to the first order polynomial

$$(F_1 \gamma + F_2)\lambda + (\gamma F_2 - F_1) = 0 \tag{92}$$

which is Hurwitz for $\gamma \varepsilon \Gamma$ provided the coefficients are positive

$$F_1 \gamma + F_2 > 0, \quad \gamma \varepsilon \Gamma \tag{93a}$$

$$\gamma F_2 - F_1 > 0, \quad \gamma \varepsilon \Gamma \tag{93b}$$

Evaluating the first inequality at $\gamma = 1$ and the second at $\gamma = -1$ give the contradicting inequalities

$$F_2 > -F_1 \tag{94a}$$

$$F_2 < -F_1 \tag{94b}$$

That is, there exists no $F = (F_1, F_2)$ satisfying (69) for $\gamma \varepsilon [-1, 1]$ which is a subset of Γ. Consequently, there is no stable sliding mode surface (91) on which a robust sliding mode controller (52) can be designed for $\gamma \varepsilon [-1, 1]$.

The requirement for the existence of a robust stabilizing linear feedback controller (80) is that (81) is Hurwitz. The characteristic polynomial of (81) is given by

$$\lambda^2 + a_1(\gamma)\lambda + a_2(\gamma) = 0 \tag{95}$$

where robustness follows from positiveness of the coefficients

$$a_1(\gamma) = K_2 + \gamma K_1 > 0, \quad \gamma \varepsilon \Gamma \tag{96a}$$

$$a_2(\gamma) = \gamma K_2 - K_1 + 1 > 0, \quad \gamma \varepsilon \Gamma \tag{96b}$$

The following gain vector $K = (K_1, K_2)$ provides a robust linear controller (80)

$$K_1 = 0 \tag{97a}$$

$$K_2 = \frac{1}{M + \varepsilon} \tag{97b}$$

where $\varepsilon > 0$. Substitution of the gain vector (97) into (96) gives

$$K_2 > 0, \quad \gamma \varepsilon \Gamma \tag{98a}$$

$$\gamma > -(M + \varepsilon), \quad \gamma \varepsilon \Gamma \tag{98b}$$

The inequalities (96) are met. Thus (81) is Hurwitz which implies that the linear controller defined by (97) robustly stabilizes (88). Consequently, (88) has a robust stabilizing linear controller but no stabilizing sliding mode controller.

6. ROBUST CONTROL IN THE ABSENCE OF MATCHING CONDITIONS: MULTI-INPUT

The multi-input case parallels that of the scalar case, Section 4. We consider a condition similar to (69) and show that it leads to necessary and sufficient conditions for the existence of a sliding mode controller (7). In this section $B(\gamma), \gamma \varepsilon \Gamma$, is an nxm uncertain matrix with full rank $(m \leq n)$ We consider system (1) for which Assumption I holds. Our main result for a robust sliding mode controller is given in the next theorem.

THEOREM 8. A robust stabilizing sliding mode controller (7) exists for system (1) in which Assumption I holds if, and only if, the following determinant is Hurwitz:

$$\det\begin{bmatrix} A(\gamma) - sI & B(\gamma) \\ F & 0 \end{bmatrix} = 0, \quad \gamma \varepsilon \Gamma \tag{99}$$

PROOF: The reciprocal form of (99) is

$$a_\gamma(s)\ \det\left[F[sI - A(\gamma)]^{-1}B(\gamma)\right] = 0, \quad \gamma \in \Gamma \tag{100}$$

where $a_\gamma(s)$, defined by (73), is the determinant of $[sI - A(\gamma)]$ which is the open-loop characteristic polynomial of $A(\gamma)$. Since Assumption I holds there is a coordinate transformation $T(\gamma)$

$$z = T(\gamma)x \tag{101}$$

which takes (1) into a controllable companion form, [34],

$$\dot{z} = A_z(a(\gamma))z + B_z u(x) \tag{102}$$

where

$$A_z(a(\gamma)) = T(\gamma)\,A(\gamma)\,T^{-1}(\gamma) \tag{103a}$$

$$B_z(\gamma) = T(\gamma)\,B(\gamma) \tag{103b}$$

The mxm matrix $B_z(\gamma)$ is formed from m columns of the nxn identity matrix. The dependence of $B_z(\gamma)$ on the uncertainty γ follows from the fact that the selection of the m columns may depend on $\gamma \epsilon \Gamma$. The nxn matrix $A_z(a(\gamma))$ is in block controllable companion form. Such companion forms are described in [32]-[34]. In view of the Transformation (101) we can rewrite (100) as

$$\det\left[\left[FT^{-1}(\gamma)\right]\ a_\gamma(s)\ T(\gamma)[sI - A(\gamma)]^{-1}B(\gamma)\right] = 0 \tag{104}$$

Consider the last two factors

$$a_\gamma(s)\ \left[T(\gamma)[sI - A(\gamma)]^{-1}B(\gamma)\right] \tag{105}$$

which in z-coordinates is given by

$$a_\gamma(s)\ \left[[sI - A_z(a(\gamma))]^{-1}B_z(\gamma)\right] \tag{106}$$

which is equivalent to

$$Adj[sI - A_z(a(\gamma))]B_z(\gamma) \tag{107}$$

where Adj is the matrix adjoint operation. Consider the definition of an nxn symmetric, positive definite matrix $P(\gamma)$ and the definition of an mxm symmetric, positive definite matrix $R(\gamma)$ such that

$$R^{-1}(\gamma)B_z^T(\gamma)P(\gamma)T(\gamma) = F, \quad \gamma \in \Gamma \tag{108}$$

That is, $P(\gamma)$ must be such that

$$R^{-1}(\gamma)B_z^T(\gamma)P(\gamma) = FT^{-1}(\gamma), \quad \gamma \in \Gamma \tag{109}$$

Furthermore, consider the Lyapunov equation

$$P(\gamma)A_z(\hat{a}(\gamma)) + A_z^T(\hat{a}(\gamma))P(\gamma) + Q(\gamma) = 0, \quad \gamma \varepsilon \Gamma \tag{110}$$

where $Q(\gamma) > 0$ and $Q(\gamma) = Q^T(\gamma), \gamma \varepsilon \Gamma$. A necessary and sufficient condition that $A_z(\hat{a}(\gamma))$ be asymptotically stable and $P(\gamma)$ be symmetric, positive definite and satisfy the constraint (109) is that the determinant of the following mxm matrix (111a) be Hurwitz and that the following mxm matrix (111b) be positive definite , [32]:

$$B_z^T(\gamma)P(\gamma)Adj[sI - A_z(\gamma)]B_z(\gamma), \quad \gamma \varepsilon \Gamma \tag{111a}$$

$$B_z^T(\gamma)P(\gamma)B_z(\gamma) > 0, \quad \gamma \varepsilon \Gamma \tag{111b}$$

From (104), (107), (109) and (111) it follows that (99) is necessary and sufficient in order that for each $\gamma \varepsilon \Gamma$ there exist a symmetric, positive definite $P(\gamma)$ satisfying (109) and a stable $A_z(\hat{a}(\gamma))$ such that the Lyapunov equation (110) is satisfied. The theorem now follows from Theorem 1. Define $\sigma(\gamma), \gamma \varepsilon \Gamma$

$$\sigma(\gamma) = B_z^T(\gamma)\left[A_z(a(\gamma)) - A_z(\hat{a}(\gamma))\right] \tag{112}$$

By the canonical form of A_z and B_z it follows that

$$A_z(\hat{a}(\gamma)) = A_z(a(\gamma)) - B_z(\gamma)\sigma(\gamma) \tag{113}$$

Define $K(\gamma), \gamma \varepsilon \Gamma$, as

$$K(\gamma) = \sigma(\gamma)T(\gamma) \tag{114}$$

Transforming (113) from z-coordinates to x-coordinates using (101) yields the following asymptotically stable matrix.

$$\overline{A}(\gamma) = A(\gamma) - B(\gamma)K(\gamma) \tag{115}$$

Thus condition (i) of Assumption II is met. Transforming (108) from z-coordinates to x-coordinates using (101) gives

$$F = R^{-1}(\gamma)B^T(\gamma)\overline{P}(\gamma) \tag{116}$$

where $\overline{P}(\gamma)$ satisfies the Lyapunov equation which is transformed from (110)

$$\overline{P}(\gamma)\overline{A}(\gamma) + \overline{A}^T(\gamma)\overline{P}(\gamma) + \overline{Q}(\gamma) = 0, \quad \gamma \varepsilon \Gamma \tag{117}$$

where

$$\overline{P}(\gamma) = T^T(\gamma)P(\gamma)T(\gamma) \tag{118a}$$

$$\overline{Q}(\gamma) = T^T(\gamma)Q(\gamma)T(\gamma) \tag{118b}$$

Thus condition (ii) of Assumption II is met. Consequently all conditions of Theorem 1 are satisfied. The existence of a robust sliding mode controller (7) now follows.

The existence of a robust stabilizing sliding mode controller implies the existence of a robust linear controller. This result is given in the next theorem which parallels the scalar result, Theorem 7:

THEOREM 9: The existence of a stabilizing sliding mode controller (7) for system (1) implies the existence of a robust stabilizing linear controller

$$u = -Kx \tag{119}$$

PROOF: A necessary and sufficient condition for the existence of a robust stabilizing linear controller is that the determinant

$$\det\begin{bmatrix} sI - A(\gamma) & B(\gamma) \\ -K & I_m \end{bmatrix} = 0, \quad \gamma \varepsilon \Gamma \tag{120}$$

is Hurwitz where I_m is the mxm identity matrix. Paralleling the developement (81) - (84) the determinant (120) is Hurwitz if, and only if, the mxm matrix

$$a_\gamma^c(s) = a_\gamma(s) \det[I_m + K[sI - A(\gamma)]^{-1} B(\gamma)], \quad \gamma \varepsilon \Gamma \tag{121}$$

is Hurwitz. If a robust stabilizing sliding mode controller (7) exists then there exists an mxn matrix F such that (99) is Hurwitz. Consequently, (100) is Hurwitz. For an arbitrary mxm matrix C define the gain matrix

$$K = CF \tag{122}$$

substitution of (122) into (121) gives

$$a_\gamma^c(s) = a_\gamma(s) \det[I_m + CF[sI - A(\gamma)]^{-1} B(\gamma)], \quad \gamma \varepsilon \Gamma \tag{123}$$

In view of (73) we can rewrite (123) as

$$a_\gamma^c(s) = \det[\lambda^n I_m + CF\, Adj[sI - A(\gamma)]\, B(\gamma) + (a_\gamma(s) - \lambda^n)\, I_m], \quad \gamma \in \Gamma \tag{124}$$

Since the mxm matrix (100) is Hurwitz, it follows that there exists an mxm matrix C with sufficiently "large elements" such that (124) is Hurwitz, [33]. The last term is dominated by the second term. The control law (119) robustly stabilizes (1) for a " sufficiently large" C matrix in (122).

7. SUMMARY

A linear time-invariant uncertain system is investigated for robust stabilization. The uncertainties belong to a compact subset of multi-dimensional Euclidean space. The dynamics and input matrices are continuous functions of uncertainty. The system is controllable for each uncertainty, Assumption I. In Assumption II two general conditions are stated which involve an uncertain Lyapunov equation. The first condition deals with the existence of an uncertain gain matrix for stabilizing the system. The second deals with the existence of a constant F matrix which has the appearance of a Riccati gain matrix. F is the product of three uncertain quantities one of which is the uncertain solution $P(\gamma)$ of the Lyapunov equation. Another is the $R(\gamma)$ matrix which is assumed in Assumption III to form a positive definite matrix when added to its transpose.

A general robustness result is established in Theorem 1. It states that a robust stabilizing sliding mode controller exists under the general Assumptions I - III. In Theorem 2 we prove that the matching conditions of uncertain systems satisfy the Assumptions I - III.

Robust control in the absence of matching conditions is examined in Theorems 3, 4 and 5 for scalar control input. For such systems necessary and sufficient conditions are given for the existence of robust stabilizing sliding mode controllers. In Theorem 4 we show that systems satisfying such conditions also meet Assumptions I - III. Theorem 5 goes one step further and shows the existence of a robust linear control for such systems. The existence of a robust sliding mode controller is shown to depend on a minimum phase condition, Theorem 6. In Section 5 we give an example of a simple system which admits a robust linear controller but no robust sliding mode controller that stabilizes the system.

In Section 6 we investigate robust control in the absence of matching conditions for multi-input systems. In Theorem 8 we show that a certain determinant being Hurwitz is necessary and sufficient for the existence of a sliding mode controller. A similar condition is stated in Theorem 9 for the existence of a robust linear controller.

REFERENCES

1. GUTMAN, S. and LEITMANN, G., "On a class of Linear Differential Games", Journal of Optimization Theory and Applications, Vol. 17, Nos 5-6, pp. 511-522, 1975.

2. LEITMANN, G., "Stabilization of Dynamical Systems under Bounded Input Disturbance and Parameter Uncertainty," Proceed. of 2nd Kingston Conference on Differential Games and Control Theory II, M. Dekker, New York, 1976.

3. GUTMAN, S. and LEITMANN, G., "Stabilizing Feedback Control for Dynamical Systems with Bounded Uncertainty," Proceedings of IEEE Conference on Decision and Control, 1976.

4. LEITMANN, G., "Guaranteed Ultimate Boundedness for a Case of Uncertain Linear Dynamical Systems," IEEE Transactions on Automatic Control, Vol. AC-23, No. 6, 1978.

5. GUTMAN, S., "Uncertain Dynamical Systems - A Lyapunov Min-Max Approach," IEEE Transactions on Automatic Control, Vol. AC-24, No. 3, pp. 437-443, June 1979.

6. LEITMANN, G., "Guaranteed Asymptotic Stability for Some Linear Systems with Bounded Uncertainties," J. Dynamic Systems, Measurement and Control, Vol. 101, No. 3, pp. 212-216, 1979.

7. LEITMANN, G., "On the Efficacy of Nonlinear Control in Uncertain Linear Systems," Dynam. Syst., Meas. Cont., Vol. 102, No. 2, pp. 95-102 1981.

8. CORLESS, M.J. and LEITMANN, G., "Continuous State Feedback Guaranteeing Uniform Ultimate Boundedness for Uncertain Dynamic Systems," IEEE Transactions on Automatic Control, Vol. AC-26, No. 5, pp. 1139-1144, October 1981.

9. BARMISH, B.R. and LEITMANN, G., "On Ultimate Boundedness Control of Uncertain Systems in the Absence of Matching Assumptions," IEEE Transactions on Automatic Control, Vol. AC-27, No. 1, pp.153-158, February 1982.

10. GUTMAN, S. and PALMOR, Z., "Properties of Min-Max Controllers in Uncertain Dynamical Systems," SIAM J. Control and Optimization, Vol. 20, No. 6, pp. 850-861, November 1982.

11. BARMISH, B.R., CORLESS, M. and LEITMANN, G., "A New Class of Stabilizing Controllers for Uncertain Dynamical Systems," SIAM J. Control and Optimization, Vol. 21, No. 2, pp. 246-255, March 1983.

12. LEITMANN, G., "Deterministic Control of Uncertain Systems," The Fourth International Conference: Mathematical Modeling in Science and Technology, Zurich, August 1983.

13. BARMISH, B.R., PETERSEN, I.R. and FEUER, A., "Linear Ultimate Boundedness Control of Uncertain Dynamical Systems," Automatica, Vol. 19, No. 5, pp. 523-532, September 1983.

14. BARMISH, B.R., "Necessary and Sufficient Conditions for Quadratic Stabilizability of an Uncertain System", Journal of Optimization Theory and Applications, Vol. 46, No. 4. pp. 399-408, August 1985.

15. PETERSEN, I.R., "Structural Stabilization of Uncertain Systems: Necessity of the Matching Condition," Proceed. of 20th Allerton Conference on Communication, Control, and Computing, 1982, also in SIAM J. Control and Opimization, Vol. 23, No. 2, pp. 286-296, March 1985.

16. GALIMIDI, Alberato R. and BARMISH, B.Ross, "The Constrained Lyaponov Problem and its Application to Robust Output Feedback Stabilization", IEEE Transactions on A.C., Vol. AC-31, No. 5, pp. 410-418, May 1986.

17. PETERSEN, I.R., "Quadratic Stabilizability of Uncertain Linear Systems: Existence of a Nonlinear Stabilizing Control Does Not Imply Existence of a Linear Stabilizing Control," IEEE Transactions on Automatic Control, Vol. AC-30, No.3, pp. 291-293, March 1985.

18. PETERSEN, I.R., "Nonlinear Versus Linear Control in the Direct Output Feedback Stabilization of Linear Systems," IEEE Transactions on Automatic Control, Vol. AC-30, No. 8, pp. 799-802, August 1985.

19. STALFORD, H.L., "Necessary and Sufficient Conditions for Matching Conditions in Uncertain Systems: Scalar Input," Proceed. of the 1987 American Control Conference, pp. 879-903, June 10-12, 1987.

20. SINGH, S.N. and COELHO, A.A.R. "Ultimate Boundedness Control of Set Points of Mismatched Uncertain Linear Systems," Int. J. Systems SCI., 1983, Vol. 14, No. 7, pp.693-710.

21. SINGH, S.N. and COELHO, A.A.R., "Nonlinear Control of Mismatched Uncertain Linear Systems and Application to Control of Aircraft," Journal of Dynamic Systems, Measurement, and Control, September 1984, Vol. 106, pp. 203-210.

22. LEITMANN, G., RYAN, E.P. and STEINBERG, A., "Feedback Control of Uncertain Systems: Robustness with respect to neglected actuator and sensor dynamics," Int. J. Control, Vol. 43, No. 4, pp. 1243-1256, 1986.

23. SCHMITENDORF, W.E. and BARMISH, B.R., "Robust Asymptotic Tracking for Linear Systems with Unknown Parameters," Automatica, Vol. 22, No. 3, pp. 355-360, 1986.

24. CHEN, Y.H. and LEITMANN, G., "Robustness of Uncertain Systems in the Absence of Matching Conditions," Int. J. Control, Vol. 45, No. 5, pp. 1527-1542, 1987.

25. SCHMITENDORF, W.E. and BARMISH, B.R., "Guaranteed Asymptotic Output Stability for Systems with Constant Disturbances", Transactions of the ASME, Vol. 109, pp. 186-189, June 1987.

26. STALFORD, H. and GARRETT, F. Jr., "Robust Nonlinear Control for High Angle-of-Attack Flight," Presented at the AIAA 25th Aerospace Sciences Meeting, Reno, Nevada, paper AIAA-87-0346, January 12-15, 1987.

27. STALFORD, H., "On Robust Control of Wing Rock Using Nonlinear Control," Proceed. 1987 American Control Conference, Minneapolis Minnesota, June 10-12, 1987.

28. STALFORD, H., "Tracking at High α Using Certain Robust Nonlinear Controllers," Proceed. AIAA Guidance, Navigation and Control Conference, Monterey, California, August 17-19, 1987.

29. STALFORD, H.L., "Robust Control of Uncertain Systems in the Absence of Matching Conditions: Scalar Input," 1987 Conference on Decision and Control, Los Angeles, California, December 8-10, 1987.

30. STALFORD, H.L. and CHAO, C.-H., "Necessary and Sufficient Condition in Lyapunov Robust Control," submitted for publication, December, 1987.

31. STALFORD, H.L. and CHAO, C.-H., "On the Robustness of Linear Stabilizing Feedback for Linear Uncertain Systems," submitted for publication, December, 1987.

32. STALFORD, H. L. and CHAO, Chien-Hsiang, "A Necessary and Sufficient Condition in Lyapunov Robust Control: Multi-Input," Submitted for possible presentation at the 27th IEEE Conference on Decision and Control, Austin, Texas, December 7-9, 1988.

33. CHAO, Chien-Hsiang and STALFORD, H. L., "On the Robustness of Linear Stabilizing Feedback Control for Linear Uncertain Systems: Multi-Input," submitted for possible presentation at the 27th IEEE Conference on Decision and Control, Austin, Texas, December 7-9, 1988.

34. KAILATH, T., "Linear Systems", Prenctice-Hall, Inc., New Jersey, 1980.

35. DAVISON, E.J. and WANG, S.H., "Properties and Calculations of Transmission Zeros of Linear Multivariable System," Control System Design by Pole-Zero Assignment , F. Fallside, Editor, Academic Press, New York, pp. 16-42, 1977.

SINGULARLY PERTURBED UNCERTAIN SYSTEMS AND DYNAMIC OUTPUT FEEDBACK CONTROL

E P Ryan and Z B Yaacob

School of Mathematical Sciences
University of Bath
Bath BA2 7AY
United Kingdom

ABSTRACT

A dynamic output feedback strategy is proposed for a class of uncertain systems. Using a singular perturbation approach, a threshold measure of "fastness" of the feedback dynamics, to ensure overall system stability, is derived. This threshold is calculable in terms of known bounds on the system uncertainties but may be conservative in practice. To circumvent this drawback and to allow for bounded uncertainties with unknown bounds, an adaptive version of the strategy is then developed.

1. Introduction

We address the problem of design of dynamic output feedback controls for a class of uncertain nonlinearly perturbed linear multivariable systems. The approach is similar in concept to that of [1], and fundamentally stems from the deterministic theory developed in, for example, [2-8] (see also bibliographies therein).

Initially considering a hypothetical output y^* for the system, a (generally unrealizable) stabilizing static output feedback control is established. This static control is then approximated by a realizable compensator (with parameter $\mu \geq 0$) which filters the true system output y. Physically, the parameter μ is a measure of "fastness" for the filter dynamics; analytically, μ plays the role of a singular perturbation parameter. Using a singular perturbation analysis akin to that of [9,10], a threshold measure μ^* of "fastness" of the compensator dynamics, to ensure overall system stability, is then derived. The threshold is explicitly calculable from known system data but corresponds to a "worst-case" value and consequently may be conservative. To counteract this inherent conservatism (and to allow for bounded uncertainties with unknown bounds) an adaptive version of the compensator is also developed by an approach which is essentially that of [11] (see also [12-16] and related work in [17-23]).

2. The system

We consider uncertain nonlinearly perturbed linear systems of the form

$$\dot{x}(t) = Ax(t) + B[u(t) + g(t,x(t),u(t))], \quad x(t) \in \mathbb{R}^n, \quad u(t) \in \mathbb{R}^m \tag{1}$$

for which the only available state information is provided by the output

$$y(t) = Cx(t), \quad y(t) \in \mathbb{R}^p, \quad m \leq p \leq n \,. \tag{2}$$

The triple (C,A,B), which defines the nominal linear system, is assumed to satisfy the following.

Assumption 1: (A,B) is a controllable pair and rank $B = m$.

Assumption 2:

There exist known integer $r \geq 1$ and known matrices $F_1, F_2, \cdots, F_r \in \mathbb{R}^{m \times p}$, such that

(i) for $i = 1,2,\cdots, r-1$, $\operatorname{im} CA^{i-1}B \subset \bigcap_{j=i+1}^{r} \ker F_j$;

moreover, the matrix $C_r := F_1C + F_2CA + \cdots + F_rCA^{r-1}$ is such that

(ii) $|C_rB| \neq 0$, and

(iii) the transmission zeros of the m-input m-output linear system (C_r, A, B) lie in $\mathbb{C}^-$ (the open left half complex plane).

Example 1: If $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$, $B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, $C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, then the above assumptions hold with $r = 2$, $F_1 = [1 \ \ 1]$ and $F_2 = [1 \ \ 0]$.

Finally, we impose some structure on the uncertain function g.

Assumption 3:

g: $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ is (i) Carathéodory, with (ii) $\|g(t,x,u)\| \leq \alpha\|x\| + \beta\|u\|$ for all (t,x,u), where α and β are known constants with $\beta < 1$, and (iii) if $r \geq 2$, then g is uniformly Lipschitz in its final argument (with known Lipschitz constant λ), i.e. (if $r \geq 2$) there exists known λ, independent of (t,x), such that, for all u and v, $\|g(t,x,u)-g(t,x,v)\| \leq \lambda\|u-v\|$.

The outline of the paper is as follows:

Firstly, the problem of designing a (dynamic) output feedback compensator for system (1,2) is addressed. This is accomplished by initially considering system (1) with hypothetical output

$$y^*(t) = C_rx(t) \tag{3}$$

where C_r is as in Assumption 2. Note that, if $r = 1$ then $y^*(t) = F_1y(t)$ and hence is realizable; however, if $r \geq 2$ then $y^*(t)$ is unavailable to the controller, hence the qualifier "hypothetical". For the system (1,3) so defined, (ii) and (iii) of Assumption 2 in essence play the role of "relative degree one" and "minimum phase" conditions on the hypothetical nominal linear system triple (C_r, A, B). Under such conditions, it is known (see, for example, [11-13]) that the zero state of system (1,3) can be rendered globally uniformly asymptotically stable by static output feedback; this is reiterated in Theorem 1. However, with the exception of the case $r = 1$, such static output feedback is unrealizable in the context of the true system (1,2). Therefore, in §3, a realizable dynamic compensator is constructed for the cases $r \geq 2$, which filters the actual output y. This filter can be interpreted as providing a realizable approximation to the static hypothetical output feedback; moreover, it is shown in Theorem 2 that global uniform asymptotic stability of the zero state of (1,2) is guaranteed provided that the filter dynamics are sufficiently fast (a calculable threshold measure of fastness is provided).

Secondly, in §4, an adaptive version of the dynamic compensator is developed, which counteracts conservatism (induced by crude estimates in the analysis) inherent in the non-adaptive filter and which also dispenses with the requirement that the uncertainty parameters α, β and λ in Assumption 3 be known (however, the assumption that $\beta < 1$ remains in force and, moreover, if $r \geq 2$ then g is assumed to depend linearly on x).

3. Stabilizing static output feedback control for hypothetical system

Let $T_1 \in \mathbb{R}^{(n-m)\times n}$ be such that $\ker T_1 = \operatorname{im} B$, then

$$T = \begin{bmatrix} T_1 \\ (C_r B)^{-1} C_r \end{bmatrix} \quad \text{with inverse} \quad T^{-1} = [S_1 \vdots B]$$

is a similarity transformation which takes system (1,3) into the form

$$\dot{\tilde{x}}(t) = A_{11}\tilde{x}(t) + A_{12}\tilde{y}(t)\ , \quad \tilde{x}(t) \in \mathbb{R}^{n-m} \tag{4a}$$

$$\dot{\tilde{y}}(t) = A_{21}\tilde{x}(t) + A_{22}\tilde{y}(t) + u(t) + \tilde{g}(t,\tilde{x}(t),\tilde{y}(t),u(t))\ , \quad \tilde{y}(t) \in \mathbb{R}^{m} \tag{4b}$$

$$\tilde{g}(t,\tilde{x},\tilde{y},u) := g(t,S_1\tilde{x}+B\tilde{y},u) \tag{4c}$$

with output

$$y^{*}(t) = (C_r B)\tilde{y}(t)\ . \tag{5}$$

Note that the eigenvalues of A_{11} coincide with the transmission zeros of (C_r, A, B); thus, by virtue of Assumption 2(iii), $\sigma(A_{11}) \subset \mathbb{C}^{-}$.

Let $P_1 > 0$ be the unique positive definite solution of the Lyapunov equation

$$P_1 A_{11} + A_{11}^T P_1 + I = 0 \tag{6}$$

then we state our first result.

Theorem 1:

Define $\kappa^{*} := \|A_{22}\| + \alpha\|B\| + \frac{1}{2}[\|P_1A_{12}+A_{21}^T\|+\alpha\|S_1\|]^2$, then, for each fixed $\hat{k} > \kappa^{*}(1-\beta)^{-1}$, the static output feedback

$$u(t) = -\hat{k}(C_r B)^{-1}y^{*}(t) = -\hat{k}\,\tilde{y}(t) \tag{7}$$

renders the zero state of the hypothetical system (1,3) globally uniformly asymptotically stable.

Proof: Let $V: (\tilde{x},\tilde{y}) \mapsto \frac{1}{2}\langle\tilde{x},P_1\tilde{x}\rangle + \frac{1}{2}\|\tilde{y}\|^2$, then a straightforward calculation reveals that, along solutions $(\tilde{x}(\cdot),\tilde{y}(\cdot))$ of (4,5,7) (equivalent to (1,3,7)), the following holds almost everywhere

$$\frac{d}{dt} V(\tilde{x}(t),\tilde{y}(t)) \leq -U(\tilde{x}(t),\tilde{y}(t))$$

where

$$U(\tilde{x},\tilde{y}) := \frac{1}{2}\left\langle \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix}, M \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \right\rangle, \quad M := \begin{bmatrix} 1 & -[\|P_1A_{12}+A_{21}^T\|+\alpha\|S_1\|] \\ -[\|P_1A_{12}+A_{21}^T\|+\alpha\|S_1\|] & 2[\hat{k}(1-\beta)-\|A_{22}\|-\alpha\|B\|] \end{bmatrix}.$$

Noting the M is positive definite, the result follows. □

In the context of the true system (1,2), if $r = 1$, then the static feedback (7) is realizable as

$$u(t) = -\hat{k}(C_rB)^{-1}F_1y(t) \tag{8}$$

whence:-

Corollary 1:

Let $\hat{k}$ be as in Theorem 1. If $r = 1$ then the static output feedback (8) renders the zero state of the true system (1,2) globally uniformly asymptotically stable.

However, in all other cases ($r \geq 2$), the feedback (7) is unrealizable for the true system (1,2); in its place, we will develop a realizable dynamic compensator in the next section.

4. Cases r ≥ 2: Stabilizing dynamic output feedback for the true system (1,2)

In view of Assumption 2(i), we note that

$$y^*(t) = C_rx(t) = F_1y(t) + F_2\dot{y}(t) + \cdots + F_ry^{(r-1)}(t)$$

which can be interpreted in the frequency domain as

$$\bar{y}^*(s) = [F_1+N(s)]\bar{y}(s) ,$$

where

$$N(s) = sF_2 + \cdots + s^{r-1}F_r$$

is physically unrealizable. Our approach is to replace $N(s)$ by a physically realizable transfer matrix (filter) of the form $H_\mu(s)N(s)$ with appropriately chosen $H_\mu(s)$. To this end, let $d_i \leq r-1$ denote the degree of the highest-degree polynomial in the ith row of $N(s)$. Let constants $a_j^i > 0$, $j=2,\cdots,d_i$, be such that

$$\pi_i(s) = s^{d_i} + a_{d_i}^is^{d_i-1} + \cdots + a_2^is + 1, \quad i = 1,2,\cdots,m$$

is Hurwitz (i.e. with all its roots lying in the open left half complex plane $\mathbb{C}^-$). For $i = 1,2,\cdots,m$, define $h_i^\mu(s)$, parameterized by $\mu > 0$, as

$$h_i^\mu(s) = \frac{1}{\pi_i(\mu s)} .$$

which, interpreted as a transfer function, has minimal realization $(c_i^T,\ \mu^{-1}A_i,\ \mu^{-1}b_i)$, where

$$A_i = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -1 & -a_2^i & -a_3^i & \cdots & -a_{d_i}^i \end{bmatrix} \in \mathbb{R}^{d_i\times d_i} , \quad b_i = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^{d_i} , \quad c_i = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \in \mathbb{R}^{d_i} .$$

We now introduce the transfer matrix

$$H_\mu(s) := \mathrm{diag}\{h_i^\mu(s)\}$$

which clearly has minimal realization $(C^*, \mu^{-1}A^*, \mu^{-1}B^*)$, where

$$A^* = \mathrm{diag}\,\{A_i\} \in \mathbb{R}^{q\times q}\ ,\ B^* = \mathrm{diag}\,\{b_i\} \in \mathbb{R}^{q\times m}\ ,\ C^* = \mathrm{diag}\,\{c_i^T\} \in \mathbb{R}^{m\times q}\ ,\ \text{with } q := \sum_{i=1}^{m} d_i\ .$$

We note, in passing, that $\sigma(A^*) \subset \mathbb{C}^-$ and that $C^*(A^*)^{-1}B^* = -I$.

Let κ^* be as in Theorem 1, then, for fixed $\hat{k} > \kappa^*(1-\beta)^{-1}$, the proposed physically realizable compensator (which filters the actual output y) for system (1,2) is parameterized by μ, and has frequency domain characterization:

$$G_\mu(s) = -\hat{k}(C_rB)^{-1}[F_1 + H_\mu(s)N(s)]\ . \tag{9}$$

For notational convenience we introduce functions φ, f_1, f_2, Δf_2 and f_3, defined as follows.

$$\varphi\colon (\tilde{x},\tilde{y},\tilde{z}) \mapsto -\hat{k}(C_rB)^{-1}\,[F_1C[S_1\tilde{x}+B\tilde{y}] + C^*\tilde{z}]$$

$$f_1\colon (\tilde{x},\tilde{y}) \mapsto A_{11}\tilde{x} + A_{12}\tilde{y}$$

$$f_2\colon (t,\tilde{x},\tilde{y}) \mapsto A_{21}\tilde{x} + A_{22}\tilde{y} - \hat{k}\tilde{y} + \tilde{g}(t,\tilde{x},\tilde{y},-\hat{k}\tilde{y})$$

$$\Delta f_2\colon (t,\tilde{x},\tilde{y},\tilde{z}) \mapsto \hat{k}\tilde{y} + \varphi(\tilde{x},\tilde{y},\tilde{z}) + \tilde{g}(t,\tilde{x},\tilde{y},\varphi(\tilde{x},\tilde{y},\tilde{z})) - \tilde{g}(t,\tilde{x},\tilde{y},-\hat{k}\tilde{y})$$

$$f_3\colon (\tilde{x},\tilde{y},\tilde{z}) \mapsto A^*\tilde{z} + B^*[\,C_rB\tilde{y}-F_1C[S_1\tilde{x}+B\tilde{y}]]\ .$$

Then it is readily verified that, in the time domain and under state transformation T, the differential equations governing the dynamic output feedback controlled system may now be expressed in the form:

$$\dot{\tilde{x}}(t) = f_1(\tilde{x}(t),\tilde{y}(t))\ ,\ \ \tilde{x}(t) \in \mathbb{R}^{n-m} \tag{10a}$$

$$\dot{\tilde{y}}(t) = f_2(t,\tilde{x}(t),\tilde{y}(t)) + \Delta f_2(t,\tilde{x}(t),\tilde{y}(t),\tilde{z}(t))\ ,\ \ \tilde{y}(t) \in \mathbb{R}^m \tag{10b}$$

$$\mu\dot{\tilde{z}}(t) = f_3(\tilde{x}(t),\tilde{y}(t),\tilde{z}(t))\ ,\ \ \tilde{z}(t) \in \mathbb{R}^q\ . \tag{10c}$$

In analysing the stability of this system, we regard μ as a singular perturbation parameter. Recalling that $C^*(A^*)^{-1}B^* = -I$, we note that system (4) with control (7) is recovered on setting $\mu = 0$ in (10); thus, in the usual terminology [9,10,24], system (4,7) may be interpreted as the reduced-order system associated with the singularly perturbed system (10). The ensuing approach is akin to that of [9,10], our objective being to determine a threshold value $\mu^* > 0$ such that, for all $\mu \in (0,\mu^*)$, the zero state of system (10) is globally uniformly asymptotically stable.

Recalling that $\sigma(A^*) \subset \mathbb{C}^-$, let $P^* > 0$ be the unique symmetric positive definite solution of the Lyapunov equation

$$P^*A^* + (A^*)^TP^* + I = 0\ . \tag{11}$$

Define $W\colon \mathbb{R}^{n-m}\times\mathbb{R}^m\times\mathbb{R}^q \to [0,\infty)$ by

$$W(\tilde{x},\tilde{y},\tilde{z}) := \tfrac{1}{2}\langle w(\tilde{x},\tilde{y},\tilde{z}),\ P^*w(\tilde{x},\tilde{y},\tilde{z})\rangle \tag{12a}$$

where

$$\begin{aligned} w(\tilde{x},\tilde{y},\tilde{z}) &:= \tilde{z} + (A^*)^{-1}B^*[\,C_rB\tilde{y}-F_1C[S_1\tilde{x}+B\tilde{y}]] \\ &= (A^*)^{-1}f_3(\tilde{x},\tilde{y},\tilde{z})\ . \end{aligned} \tag{12b}$$

We now establish some preliminary lemmas:

The first is implicit in the proof of Theorem 1.

Lemma 1:

$\langle \nabla_{\tilde x} V(\tilde x,\tilde y), f_1(\tilde x,\tilde y)\rangle + \langle \nabla_{\tilde y} V(\tilde x,\tilde y), f_2(t,\tilde x,\tilde y)\rangle \leq -\alpha_0 V(\tilde x,\tilde y)$ where $\alpha_0 := [\|M^{-1}\|[\|P_1\|+1]]^{-1} > 0.$

Lemma 2: $\langle \nabla_{\tilde z} W(\tilde x,\tilde y,\tilde z), f_3(\tilde x,\tilde y,\tilde z)\rangle \leq -\beta_0 W(\tilde x,\tilde y,\tilde z)$ where $\beta_0 := \|P^*\|^{-1} > 0$.

Proof:

$$\begin{aligned}\langle \nabla_{\tilde z} W(\tilde x,\tilde y,\tilde z),\ f_3(\tilde x,\tilde y,\tilde z)\rangle &= \langle P^* w(\tilde x,\tilde y,\tilde z),\ f_3(\tilde x,\tilde y, zy)\rangle \\ &= \langle P^* w(\tilde x,\tilde y,\tilde z),\ A^* w(\tilde x,\tilde y,\tilde z)\rangle \\ &= -\tfrac{1}{2}\|w(\tilde x,\tilde y,\tilde z)\|^2 \\ &\leq -\|P^*\|^{-1} W(\tilde x,\tilde y,\tilde z)\ . \quad \Box\end{aligned}$$

Clearly, the function $\|f_1\|$ is bounded above by a calculable scalar multiple of the function $V^{\frac{1}{2}}$. In view of Assumption 3(ii), $\|f_2\|$ is also bounded above by a calculable scalar multiple of $V^{\frac{1}{2}}$. By Assumption 3(iii), $\tilde g$ is uniformly Lipschitz in its final argument (with known Lipschitz constant λ); hence,

$$\|\Delta f_2(t,\tilde x,\tilde y,\tilde z)\| \leq (1+\lambda)\|\hat k\,\tilde y + \varphi(\tilde x,\tilde y,\tilde z)\| \quad \text{for all } (t,\tilde x,\tilde y,\tilde z)$$

and, since $\hat k\,\tilde y + \varphi(\tilde x,\tilde y,\tilde z) = -\hat k (C,B)^{-1} C^* w(\tilde x,\tilde y,\tilde z)$, it follows that $\|\Delta f_2\|$ is bounded above by a calculable scalar multiple of $W^{\frac{1}{2}}$. Therefore, we may conclude:

Lemma 3:

There exist calculable constants θ_0, ψ_1, ψ_2 and η_0 such that, for all $(t,\tilde x,\tilde y,\tilde z)$,

(i) $\langle \nabla_{\tilde x} W(\tilde x,\tilde y,\tilde z),\ f_1(\tilde x,\tilde y)\rangle \leq \theta_0 V^{\frac{1}{2}}(\tilde x,\tilde y) W^{\frac{1}{2}}(\tilde x,\tilde y,\tilde z)$,

(ii) $\langle \nabla_{\tilde y} W(\tilde x,\tilde y,\tilde z),\ f_2(t,\tilde x,\tilde y) + \Delta f_2(t,\tilde x,\tilde y,\tilde z)\rangle \leq \psi_1 W(\tilde x,\tilde y,\tilde z) + \psi_2 V^{\frac{1}{2}}(\tilde x,\tilde y) W^{\frac{1}{2}}(\tilde x,\tilde y,\tilde z)$,

(iii) $\langle \nabla_{\tilde y} V(\tilde x,\tilde y),\ \Delta f_2(t,\tilde x,\tilde y,\tilde z)\rangle \leq \eta_0 V^{\frac{1}{2}}(\tilde x,\tilde y) W^{\frac{1}{2}}(\tilde x,\tilde y,\tilde z)$.

The next theorem demonstrates that system (10) is asymptotically stable for all $\mu > 0$ sufficiently small.

Theorem 2:

Let κ^* be as in Theorem 1 and define $\mu^* := \alpha_0 \beta_0 [\alpha_0 \psi_1 + \eta_0(\theta_0 + \psi_2)]^{-1} > 0$. Then, for each fixed $\hat k > \kappa^*(1-\beta)^{-1}$ and fixed $\mu \in (0,\mu^*)$, the zero state of system (10) is globally uniformly asymptotically stable.

Proof: Define the positive definite quadratic form (Lyapunov function candidate) $\mathcal{W}$ by

$$\mathcal{W}(\tilde x,\tilde y,\tilde z) := V(\tilde x,\tilde y) + [\theta_0 + \psi_2]^{-1} \eta_0 W(\tilde x,\tilde y,\tilde z)$$

then, invoking Lemmas 1, 2 and 3, the following holds almost everywhere along solutions $(\tilde x(\cdot),\tilde y(\cdot),\tilde z(\cdot))$ of (10):

$$\frac{d}{dt}\,\mathcal{W}(\tilde x(t),\tilde y(t),\tilde z(t)) \leq -\left\langle \begin{bmatrix} V^{\frac{1}{2}}(\tilde x(t),\tilde y(t)) \\ W^{\frac{1}{2}}(\tilde x(t),\tilde y(t),\tilde z(t)) \end{bmatrix},\ \mathcal{M} \begin{bmatrix} V^{\frac{1}{2}}(\tilde x(t),\tilde y(t)) \\ W^{\frac{1}{2}}(\tilde x(t),\tilde y(t),\tilde z(t)) \end{bmatrix} \right\rangle$$

where

$$\mathcal{M} := \begin{bmatrix} \alpha_0 & -\eta_0 \\ -\eta_0 & (\mu^{-1}\beta_0 - \psi_1)(\theta_0 + \psi_2)^{-1}\eta_0 \end{bmatrix} .$$

Noting that $\mathcal{M}$ is positive definite, the result follows. □

In summary, let $\Sigma_\mu = (\mathcal{F},\ \mu^{-1}\mathcal{A},\ \mu^{-1}\mathcal{B})$ realize (minimally) the component $H_\mu(s)N(s)$ of the proposed compensator (9), then the overall controlled system has the structure shown in Figure 1.

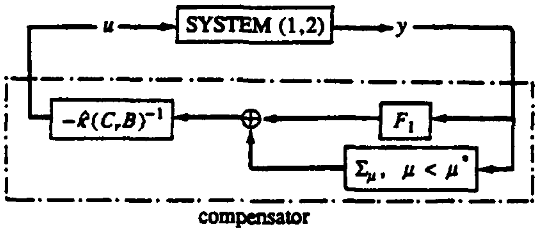


Figure 1

The governing equations (equivalent to (10)) can be expressed as

$$\dot{x}(t) = Ax(t) + B[u(t) + g(t,x(t),u(t))],\ \ x(t) \in \mathbb{R}^n \tag{13a}$$

$$\mu\dot{z}(t) = \mathcal{A}z(t) + \mathcal{B}y(t),\ \ z(t) \in \mathbb{R}^q,\ \ \mu < \mu^*, \tag{13b}$$

$$y(t) = Cx(t) \in \mathbb{R}^p \tag{13c}$$

$$u(t) = -\hat{k}(C,B)^{-1}[F_1y(t) + \mathcal{F}z(t)] \in \mathbb{R}^m\ ,\ \ \hat{k} > \kappa^*(1-\beta)^{-1}\ . \tag{13d}$$

Clearly, the *threshold values* κ^* *and* μ^* *are central to this design. Since these* values are determined via a "worst-case" analysis, it is to be expected that, in practice, the compensator will be conservative. In the next section, a stabilizing *adaptive* version of the compensator is developed; however, in the case $r \geq 2$, this is achieved at the expense of imposing further structure on the uncertain function g.

5. Adaptive compensator

5.1 Case I: $r = 1$

If Assumption 2 holds with $r = 1$ then, by Corollary 1, system (1,2) is asymptotically stabilized by the static output feedback (8) with $\hat{k} > \kappa^*(1-\beta)^{-1}$ provided, of course, that F_1 and C,B are known and that sufficient *a priori* information is avilable to compute the (conservative) gain threshold $\kappa^*(1-\beta)^{-1}$. We now consider the case for which the latter information is unavailable, i.e. we only assume knowledge of F_1 and C,B and, in particular, the constants α and $\beta < 1$ in Assumption 3 may be unknown. All other assumptions remain in force.

Replace fixed $\hat{\kappa}$ in (8) by variable $\kappa(t)$ to yield

$$u(t) = -\kappa(t)(C_rB)^{-1}F_1y(t) \tag{14a}$$

and let $\kappa(t)$ evolve according to the adaptation law

$$\dot{\kappa}(t) = \|(C_rB)^{-1}F_1y(t)\|^2 \ . \tag{14b}$$

then:-

Theorem 3:

For all initial data $(t_0, x(t_0), \kappa(t_0)) \in \mathbb{R}\times\mathbb{R}^n\times[0,\infty)$, the adaptively controlled system (1,2,14) exhibits the following properties:

(i) $\lim\limits_{t\to\infty} \kappa(t)$ exists and is finite;

(ii) $\lim\limits_{t\to\infty} \|x(t)\| = 0.$

Proof: For fixed (but unknown) $\hat{\kappa} > \kappa^*(1-\beta)^{-1}$ and under the similarity transformation T, system (1,2,14) may be expressed as

$$\dot{\tilde{x}}(t) = A_{11}\tilde{x}(t) + A_{12}\tilde{y}(t) \tag{15a}$$

$$\dot{\tilde{y}}(t) = A_{21}\tilde{x}(t) + A_{22}\tilde{y}(t) - \hat{\kappa}\tilde{y}(t) - [\kappa(t)-\hat{\kappa}]\tilde{y}(t) + \tilde{g}(t,\tilde{x}(t),\tilde{y}(t),-\kappa(t)\tilde{y}(t)) \tag{15b}$$

$$\dot{\kappa}(t) = \|\tilde{y}(t)\|^2 \ . \tag{15c}$$

Let U and V be as in the proof of Theorem 1 and define the positive definite (since $\beta < 1$) function

$$\mathcal{V}\colon (\tilde{x},\tilde{y},\kappa) \mapsto V(\tilde{x},\tilde{y}) + \tfrac{1}{2}(\kappa-\hat{\kappa})^2 - \tfrac{1}{2}\beta(\kappa-\hat{\kappa})\,|\kappa-\hat{\kappa}|\ .$$

Then, along solutions $(\tilde{x}(\cdot),\tilde{y}(\cdot),\kappa(\cdot))$ of (15), the following holds almost everywhere

$$\begin{aligned}\frac{d}{dt}\mathcal{V}(\tilde{x}(t),\tilde{y}(t),\kappa(t)) &\le -U(\tilde{x}(t),\tilde{y}(t)) - \beta\hat{\kappa}\|\tilde{y}(t)\|^2 - (\kappa(t)-\hat{\kappa})\|\tilde{y}(t)\|^2 + \beta\kappa(t)\|\tilde{y}(t)\|^2 \\ &\qquad + [(\kappa(t)-\hat{\kappa}) - \beta\,|\kappa(t)-\hat{\kappa}|\,]\|\tilde{y}(t)\|^2 \\ &\le -U(\tilde{x}(t),\tilde{y}(t)) \ . \end{aligned} \tag{16}$$

Since U is positive definite, we conclude that $t \mapsto (\tilde{x}(t),\tilde{y}(t),\kappa(t))$ is bounded and since $t \mapsto \kappa(t)$ is also monotonic, assertion (i) of the theorem follows. Furthermore, in view of (16), we have $\int_{t_0}^{\infty} U(\tilde{x}(t),\tilde{y}(t))dt \le \mathcal{V}(\tilde{x}(t_0,\tilde{y}(t_0),\kappa(t_0)) < \infty$ and hence, since U and V are positive definite quadratic forms, $\int_{t_0}^{\infty} V(\tilde{x}(t),\tilde{y}(t))dt < \infty$; moreover, $\dot{V}(\tilde{x}(\cdot),\tilde{y}(\cdot))$ is essentially bounded from above. Therefore, we conclude that $V(\tilde{x}(t),\tilde{y}(t)) \to 0$ as $t \to \infty$ (see Lemma 6.3 of [22]), whence assertion (ii) of the theorem. □

5.2 Case II: $r \ge 2$

Before describing the adaptive strategy in this case, it is remarked that the argument used in establishing Theorem 3 cannot be carried over directly. Instead, we will base our approach on that of Mårtensson [11]. For this reason, further conditions are imposed on the uncertain function g. In particular, Assumption 3 is now replaced by:

Assumption 3′:
There exist a bounded continuous function $\Delta A: \mathbb{R} \to \mathbb{R}^{m\times n}$, a Carathéodory function $g_a: \mathbb{R}\times\mathbb{R}^m \to \mathbb{R}^m$ which is uniformly Lipschitz in its second argument, and a constant $\beta < 1$ such that

(i) $g(t,x,u) = \Delta A(t)x + g_a(t,u)$, for all (t,x,u),

(ii) $g_a(t,u) \leq \beta\|u\|$, for all (t,u),

and

(iii) $(C, A+B\Delta A(\cdot))$ is uniformly completely observable in the sense of [25].

Note that, if Assumption 3′ holds, then Assumption 3 holds *a fortiori* with $\alpha = \sup_t \|\Delta A(t)\|$ provided that α, β and the Lipschitz constant for $g_a(t,\cdot)$ are known. However, knowledge of these constants is *not* required here.

Example 2: With (C,A,B) defined as in Example 1 of §2, Assumption 3′(i) holds for any bounded continuous $\Delta A: t \mapsto (\Delta a_1(t), \Delta a_2(t), \Delta a_3(t))$.

Now replace fixed $\hat{k}$ in (13d) by variable $\kappa(t) > 0$ and replace fixed μ in (13b) by $(\delta\kappa(t))^{-1}$, where $\delta > 0$ is a constant (design parameter) and let $\kappa(t)$ *evolve according* to the adaptation law (other adaptation laws may be feasible, as discussed in [20])

$$\dot{\kappa}(t) = \|y(t)\|^2 + \|z(t)\|^2 .$$

Writing (as in [11])

$$x^\dagger(t) = \begin{bmatrix} x(t) \\ z(t) \end{bmatrix}, \quad u^\dagger(t) = \begin{bmatrix} u(t) \\ \dot{z}(t) \end{bmatrix}, \quad y^\dagger(t) = \begin{bmatrix} y(t) \\ z(t) \end{bmatrix},$$

then the overall adaptively controlled system may be expressed in the form

$$\dot{x}^\dagger(t) = A^\dagger(t)x^\dagger(t) + B^\dagger[u^\dagger(t) + g^\dagger(t,u^\dagger(t))] , \quad x^\dagger(t) \in \mathbb{R}^{n+q} , \tag{17a}$$

$$y^\dagger(t) = C^\dagger x^\dagger(t) \in \mathbb{R}^{p+q} , \tag{17b}$$

$$u^\dagger(t) = -\kappa(t)K^\dagger y^\dagger(t) \in \mathbb{R}^{m+q} , \tag{17c}$$

$$\dot{\kappa}(t) = \|y^\dagger(t)\|^2 , \tag{17d}$$

where

$$A^\dagger(t) := \begin{bmatrix} A+B\Delta A(t) & 0 \\ 0 & 0 \end{bmatrix}, \quad B^\dagger := \begin{bmatrix} B & 0 \\ 0 & I \end{bmatrix}, \quad C^\dagger := \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}, \tag{17e}$$

and

$$K^\dagger := \begin{bmatrix} (C,B)^{-1}F_1 & (C,B)^{-1}\mathcal{F} \\ -\delta\mathcal{B} & -\delta\mathcal{A} \end{bmatrix}, \quad g^\dagger(t,u^\dagger) := \begin{bmatrix} g_a(t,u) \\ 0 \end{bmatrix}. \tag{17f}$$

The stability of system (17) will now be investigated. We first require the following lemma (essentially a non-autonomous version of Mårtensson's lemma [11]).

Lemma 4:

Let $x^\dagger : \mathbb{R} \to \mathbb{R}^{n+q}$ satisfy

$$\dot{x}^\dagger(t) = A^\dagger(t)x^\dagger(t) + B^\dagger[v(t) + g^\dagger(t,v(t))]$$

where $v: \mathbb{R} \to \mathbb{R}^{m+q}$ is measurable. Then, for each fixed $\tau > 0$ there exists a constant $c > 0$ such that, for all t,

$$\|x^\dagger(t)\|^2 \le c \int_{t-\tau}^{t} [\|y^\dagger(s)\|^2 + \|v(t)\|^2] \, ds \ .$$

Proof: Let $\Phi(\cdot,\cdot)$ denote the state transition matrix function generated by $A+B\Delta A(\cdot)$ and define the observability Gramian for the pair $(C,A+B\Delta A(\cdot))$ in the usual manner, that is,

$$\Gamma(t,s) := \int_s^t \Phi^T(\sigma,s)C^T C\Phi(\sigma,s) \, d\sigma \ .$$

Now, for some constants k_1 and ω, we have $\|\exp At\| \le k_1 e^{\omega t}$ and, since $\Delta A(\cdot)$ is bounded (by assumption), there exists constant k_2 such that $\|B\Delta A(t)\| \le k_2$. By standard perturbation theory, we conclude that

$$\|\Phi(t,s)\| \le k_1 e^{(\omega+k_1k_2)(t-s)} \quad \text{for all } t,s \ .$$

Clearly, the state transition matrix function $\Phi^\dagger(\cdot,\cdot)$ generated by $A^\dagger(\cdot)$ is given by

$$\Phi^\dagger(t,s) = \begin{bmatrix} \Phi(t,s) & 0 \\ 0 & I \end{bmatrix} ,$$

whence

$$\|\Phi^\dagger(t,s)\| \le c_1(t-s) \quad \text{for all } t,s, \tag{18a}$$

where

$$c_1 : \sigma \mapsto 1 + k_1 e^{(\omega+k_1k_2)\sigma} \ . \tag{18b}$$

The observability Gramian for the pair $(C^\dagger, A^\dagger(\cdot))$ is given by

$$\Gamma^\dagger(t,s) := \begin{bmatrix} \Gamma(t,s) & 0 \\ 0 & (t-s)I \end{bmatrix} ,$$

and, since $(C,A+B\Delta A(\cdot))$ is uniformly completely observable (by assumption), we may conclude (see [25]) that, for each fixed $\tau \ge 0$, there exist positive constants c_2 and c_3 such that, for all t,

$$c_2\|\zeta\|^2 \le \langle \zeta \ , \ \Gamma^\dagger(t,t-\tau)\zeta \rangle \le c_3\|\zeta\|^2 \quad \forall \ \zeta \in \mathbb{R}^{n+q} \ . \tag{19}$$

Now define the measurable function $v^\dagger : t \mapsto v(t)+g^\dagger(t,v(t))$ and note that $\|v^\dagger(t)\| \le (1+\beta)\|v(t)\|$. Then,

$$x^\dagger(t) = \Phi^\dagger(t,t-\tau)x^\dagger(t-\tau) + \int_{t-\tau}^{t} \Phi^\dagger(t,s)B^\dagger v^\dagger(s) \, ds$$

whence

$$\begin{aligned} \|x^\dagger(t)\|^2 &\le 2\|\Phi^\dagger(t,t-\tau)x^\dagger(t-\tau)\|^2 + 2\|\int_{t-\tau}^{t} \Phi^\dagger(t,s)B^\dagger v^\dagger(s) \, ds\|^2 \\ &\le 2c_4\|x^\dagger(t-\tau)\|^2 + 2c_5(1+\beta)^2\|B^\dagger\|^2 \int_{t-\tau}^{t} \|v(s)\|^2 ds \ , \end{aligned} \tag{20a}$$

wherein (18) has been used, and

$$c_4 := c_1^2(\tau), \quad c_5 := \int_0^\tau c_1^2(s)\, ds\ . \tag{20b}$$

Also, invoking both (18) and (19),

$$\begin{aligned}\|x^\dagger(t-\tau)\|^2 &\le c_2^{-1}\ \langle x^\dagger(t-\tau), \Gamma^\dagger(t,t-\tau)x^\dagger(t-\tau)\rangle \\ &= c_2^{-1}\int_{t-\tau}^t \|\ y^\dagger(s) - C^\dagger\int_{t-\tau}^s \Phi^\dagger(s,\sigma)B^\dagger v^\dagger(\sigma)\ d\sigma\ \|^2\ ds \\ &\le 2c_2^{-1}\left[\ \int_{t-\tau}^t \|y^\dagger(s)\|^2 ds + c_6\tau(1+\beta)^2\|C^\dagger\|^2\|B^\dagger\|^2\int_{t-\tau}^t \|v(s)\|^2 ds\right]\ ,\end{aligned} \tag{21a}$$

where

$$c_6 := \int_0^\tau\int_0^s c_1^2(\sigma) d\sigma\ ds\ . \tag{21b}$$

Combining (20) and (21) yields the required result. □

Theorem 4:

For all initial data $(t_0, x^\dagger(t_0), \kappa(t_0)) \in \mathbb{R}\times\mathbb{R}^{n+q}\times(0,\infty)$, system (17) exhibits the following properties:

(i) $\lim_{t\to\infty} \kappa(t)$ exists and is finite;

(ii) $\lim_{t\to\infty} \|x^\dagger(t)\| = 0$.

Proof: Seeking a contradiction to (i), suppose that the monotonically increasing function $t \mapsto \kappa(t)$ is unbounded. Then, for some $t_1 \in [0,\infty)$, $\kappa(t_0+t_1) = \hat{\kappa} > \kappa^*(1-\beta)^{-1}$ and $(\delta\kappa(t_0+t_1))^{-1} = \mu < \mu^*$. Now, an argument similar to that used in the proof of Theorem 2 can be adopted to establish the existence of a positive definite quadratic form $x^\dagger \mapsto \mathcal{V}^\dagger(x^\dagger)$ and positive constant ρ such that the following holds on solutions $(x^\dagger(\cdot), \kappa(\cdot))\colon [t_0,\infty) \to \mathbb{R}^{n+q}\times(0,\infty)$ of (17):

$$\frac{d}{dt}\ \mathcal{V}^\dagger(x^\dagger(t)) \le -\rho\ \mathcal{V}^\dagger(x^\dagger(t)) \ \text{ for almost all }\ t \ge t_0 + t_1\ .$$

Thus $x^\dagger\colon [t_0,\infty) \to \mathbb{R}^{n+q}$ ultimately tends exponentially to zero; hence, both $x^\dagger$ and $y^\dagger$ are square integrable on $[t_0,\infty)$, which, in view of (17d), contradicts our supposition that the function κ is unbounded. This establishes assertion (i) of the theorem.

It remains to show that $x^\dagger(t) \to 0$ as $t \to \infty$. Clearly, (i) ensures that $y^\dagger$ is square integrable on $[t_0,\infty)$ and, in view of (17c), that $u^\dagger$ is a bounded linear transformation of $y^\dagger$. Thus, we may conclude that $u^\dagger$ is also square integrable. Now, by Lemma 4, we have

$$\begin{aligned}\|x^\dagger(t)\|^2 &\le c\int_{t-\tau}^t [\|y^\dagger(s)\|^2 + \|u^\dagger(s)\|^2]\ ds \\ &= c\int_{t_0}^t [\|y^\dagger(s)\|^2 + \|u^\dagger(s)\|^2]\ ds\ -\ c\int_{t_0}^{t-\tau} [\|y^\dagger(s)\|^2 + \|u^\dagger(s)\|^2]\ ds\ .\end{aligned}$$

Therefore, $\|x^\dagger(t)\| \to 0$ as $t \to \infty$. □

6. Discontinuous feedback

In this final section, some possible generalizations of the proposed compensators are briefly discussed. In [23] and for the case $r = 1$ only, a wider class of uncertain functions g is studied; specifically, Assumption 3 (ii) is replaced by the condition

$$\|g(t,x,u)\| \leq \alpha\|x\| + \beta\|u\| + \gamma\xi(Cx) \quad \text{for all } (t,x,u)$$

with α and $\beta < 1$ as before and where γ is a constant (assumed known in the non-adaptive case) and ξ is a known continuous function. Thus, loosely speaking, in [23] a non-cone-bounded component of uncertainty is allowed but this is required to be bounded by a function of the system output y. In the context of this more general class of systems, the assertion of Corollary 1 of the present paper remains true for fixed $\hat{k} > (1-\beta)^{-1}\max\{\kappa^*,\gamma\}$ if (8) is replaced by the generalized feedback

$$u(t) \in -\hat{k}\left[(C_rB)^{-1}F_1y(t) + \xi(y(t))\mathcal{N}(y(t))\right], \tag{22a}$$

where the set-valued map $y \mapsto \mathcal{N}(y) \subset \mathbb{R}^m$ in essence models a discontinuous control component and is given by

$$\mathcal{N}(y) := \begin{cases} \{\|(C_rB)^{-1}F_1y\|^{-1}(C_rB)^{-1}F_1y\}; & F_1y \neq 0 \\ \{v:\ \|v\| \leq 1\}; & F_1y = 0, \end{cases} \tag{22b}$$

and the overall controlled system is consequently interpreted in the generalized sense of a controlled differential inclusion [26]. Furthermore, the assertions of Theorem 3 of the present paper remain true if (22) is replaced by the adaptive control

$$u(t) \in -\kappa(t)\left[(C_rB)^{-1}F_1y + \xi(y(t))\mathcal{N}(y(t))\right]$$

where $\kappa(t)$ evolves according to (14b).

In the cases $r \geq 2$, preliminary investigations indicate that again a non-cone-bounded component of uncertainty (although considerably less general than that of the preceding paragraph) can be tolerated in g and counteracted by augmenting the compensator (13d) (or its adaptive counterpart implicit in (17c,d)) with an appropriately chosen set-valued map (again essentially modelling a discontinuous control component). However, the requisite structural conditions on the non-cone-bounded uncertainty are, as might be expected, of a rather restrictive and technical nature (akin to those in [10]) and are not detailed here.

7. References

[1] A. Steinberg and E.P. Ryan, *Dynamic output feedback control of a class of uncertain systems*, IEEE Trans. Autom. Control, AC-31 (1986), pp. 1163-1165.

[2] S. Gutman, *Uncertain dynamical systems – A Lyapunov min–max approach*, IEEE Trans. Autom. Control, AC–24 (1979), pp. 437-443.

[3] G. Leitmann, *Deterministic control of uncertain systems*, Astronautica Acta, 7 (1980), pp. 1457-1461.

[4] G. Leitmann, *On the efficacy of nonlinear control in uncertain linear systems*, J. Dynamic Systems Meas. Control, 103 (1981), pp. 95-102.

[5] M. Corless and G. Leitmann, *Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems*, IEEE Trans. Autom. Control, AC-26 (1981), pp. 1139-1144.

[6] B.R. Barmish and G. Leitmann, *On ultimate boundedness control of uncertain systems in the absence of matching conditions*, IEEE Trans. Autom. Control, AC−27 (1982), pp. 153-158.

[7] B.R. Barmish, M. Corless and G. Leitmann, *A new class of stabilizing controllers for uncertain dynamical systems*, SIAM J. Control & Optimization, 21 (1983), pp. 246-255.

[8] E.P. Ryan and M. Corless, *Ultimate boundedness and asymptotic stability of a class of uncertain dynamical systems via continuous and discontinuous feedback control*, IMA J. Math. Control & Info., 1 (1984), pp. 223-242.

[9] A. Saberi and H.K. Khalil, *Quadratic−type Lyapunov functions* for *singularly perturbed systems*, IEEE Trans. Autom. Control, AC−29 (1984), pp. 542-550.

[10] M. Corless, G. Leitmann and E.P. Ryan, *Control of uncertain systems with neglected dynamics*, preprint (1988).

[11] B. Mårtensson, *The order of any stabilizing regulator is sufficient a priori information for adaptive stabilization*, Systems & Control Letters 6 (1985), pp. 87-91.

[12] C.I. Byrnes and A. Isidori, *A frequency domain philosophy for nonlinear systems, with applications to stabilization and to adaptive control*, Proc. 23rd IEEE Conf. on Decision & Control, Las Vegas (1984), pp. 1569-1573.

[13] C.I. Byrnes and J.C. Willems, *Adaptive stabilization of multivariable linear systems*, Proc. 23rd IEEE Conf. on Decision & Control, Las Vegas (1984), pp. 1574-1577.

[14] A.S. Morse, *A three-dimensional universal controller for the adaptive stabilization of any strictly proper minimum-phase system with relative degree not exceeding two*, IEEE Trans. Autom. Control, AC-30 (1985), pp. 1188-1191.

[15] D.R. Mudgett and A.S. Morse, *Adaptive stabilization of linear systems with unknown high frequency gains*, IEEE Trans. Autom. Control, AC-30 (1985), pp. 549-554.

[16] R.D. Nussbaum, *Some remarks on a conjecture in parameter adaptive control*, Systems & Control Letters 3 (1983), pp. 243-246.

[17] M. Fu and B. Ross Barmish, *Adaptive stabilization of linear systems via switching control*, IEEE Trans. Autom. Control, AC-31 (1986), pp. 1097-1103.

[18] P. Ioannou, *Adaptive stabilization of not necessarily minimum phase plants*, Systems & Control Letters 7 (1986), pp. 281-287.

[19] H. Khalil and A. Saberi, *Adaptive stabilization of a class of nonlinear systems using high-gain feedback*, IEEE Trans. Autom. Control, AC-32 (1987), pp. 1031-1035.

[20] A. Ilchmann, D.H. Owens and D. Pfatzel-Wolters, *High-gain robust adaptive controllers for multivariable systems*, Systems & Control Letters, 8 (1987), pp. 397-404.

[21] M. Corless and G. Leitmann, *Adaptive control of systems containing uncertain functions and unknown functions with uncertain bounds*, JOTA 41 (1983), pp. 155-168.

[22] M. Corless and G. Leitmann, *Adaptive control for uncertain dynamical systems*, in Dynamical Systems and Microphysics (eds: A Blaquiere & G Leitmann) (Academic Press, New York, 1984).

[23] E.P. Ryan, *Adaptive stabilization of a class of uncertain nonlinear systems: A differential inclusion approach*, Systems & Control Letters, 10 (1988), pp. 95-101.

[24] P.V. Kokotovic, R.E. O'Malley Jr. and P. Sannuti, *Singular perturbations and order reduction in control theory - an overview*, Automatica, 12 (1976), pp. 123-132.

[25] B.D.O. Anderson, *Exponential stability of linear equations arising in adaptive identification*, IEEE Trans. Autom. Control, AC-22 (1977), pp. 83-88.

[26] J.P. Aubin and A. Cellina, *Differential Inclusions*, (Springer-Verlag, New York, 1984).

A NEW APPROACH TO THE MODELLING UNCERTAINTY PROBLEM

David Bensoussan
Ecole de technologie supérieure
Université du Québec
4750, rue Henri-Julien
Case postale 1000, Succursale E
Montréal, Québec
H2T 1R0

ABSTRACT

Modelling of systems is generally done by frequency response methods or state variable methods. It is our object to show how frequency domain robustness results can be extrapolated to their state space counterpart. Using properties of input-output relations of systems, and different compatible norms, it will be shown how a corresponding frequency response robustness results can be applied. The method can be used to solve a certain class of non linear equations. It can also apply to the control of non linear uncertain multivariable systems in order to better stability, sensitivity as well as decentralized control results. It can also apply to assess the state feedback compensator, the observer and the output feedback compensation with regard to the robustness problem.

Multivariable control theory evolved in the sixties, using the state variable approach. This approach together with growing computer technology gave rise to tremendous research. Interesting results on system stability, controllability, observability, reachability and detectability were developed. This was a sharp contrast to the single input-single output frequency response approach involving polynomial approaches, Nyquist criterium, and root locus methods.

However, many of the answers given by state space methods lack the suppleness of multivariable methods as they apply to well defined models with no modelling uncertainty. Adaptive control is a partial response for the modelling uncertainty problem as far as parametric uncertainty is concerned. Clearly, in any state space representation (A, B, C, D), there is no way to predict the behaviour of eigenvalues whenever the matrix representation is modified to $(A+\Delta A,B,C,D)$. On the other hand, frequency response methods apply better to the uncertainty problem: in the case of a single input single output Nyquist diagram for instance, a Nyquist plot could be replaced by some Nyquist band representing the modelling uncertainty at each frequency.

Multivariable frequency response methods such as the inverse Nyquist area [1] multivariable Nyquist criterium [2] and multivariable root locus [3] are concerned mainly with system stability. However, the input output approach to systems [4,5,6,7,8] which apply to any normed algebraic representation of systems fit particularily to the frequency response setting. Such an approach allows us to handle the problem of modelling uncertainty. It is our purpose to show how multivariable frequency response uncertainty methods can be extrapolated to the multivariable state space uncertain models case.

It is our aim to show how these input output robustness results can be implemented in systems described by their state space form. Given a state space model with a state feedback compensation, observer output feedback compensation, we shall derive the best possible bounds on the closed loop perturbations due to some uncertainty ΔA in the dynamics of a system. Conversely, any frequency response robustness result will be shown to hold for a corresponding state space disturbed model $A + \Delta A$ and bounds on acceptable uncertainties ΔA will be deduced.

ASYMPTOTIC LINEARIZATION OF UNCERTAIN MULTIVARIABLE SYSTEMS BY SLIDING MODES

G. Bartolini
Dipartimento di Informatica, Sistemistica e Telematica
Via Opera Pia 11a - Genova (Italy)

T. Zolezzi
Dipartimento di Matematica
Via L.B.Alberti 4 - 16132 GENOVA (Italy)

A MODEL PROBLEM. We consider control systems with deterministically uncertain dynamics described by the differential inclusions

$$(1)\begin{cases} \ddot{x} \in G_1(t,x,\dot{x},y,\dot{y},u_1,u_2) \\ \ddot{y} \in G_2(t,x,\dot{x},y,\dot{y},u_1,u_2) \end{cases} \quad , t > 0.$$

Here x,y are scalar state variables and u_1,u_2 are scalar control variables, constrained by

$$(u_1,u_2) \in U$$

a given closed subset of R^2. Motivations for considering system (1) come from robotics, since some dynamic equations of kinematic chains appearing in robotics may be reduced to the form (1). See e.g. [6] for a recent treatment.

The multifunction which describes the unknown system dynamics is given by

$$(G_1, G_2)' : [0, +\infty) \times R^4 \times U \rightrightarrows R^2.$$

We assume explicit knowledge of some upper and lower bounds of the dynamics involved. Therefore Carathèodory functions

$$g_i^+ , g_i^- , i = 1,2$$

are known such that

$$G_1 = [g_1^- , g_1^+] , G_2 = [g_2^- , g_2^+] .$$

The initial state is uncertain but bounded by some known constant.

We consider a given linear time invariant (known) model

$$(2) \quad \ddot{w} = a_1\dot{w} + a_2 w \qquad + a_3 v_1 , \quad \ddot{z} = b_1\dot{z} + b_2 z + b_3 v_2$$

with scalar control variables v_1, v_2, state variables w, z, $w(0)$ and $z(0)$ fixed, and arbitrarily fixed coefficients a_i, b_i $(i = 1,\dots,3)$.

The problem we consider is to find a state feedback control law (possibly depending on instantaneous values of v_1, v_2, w, $\dot{w}$ and $z,\dot{z}$) under which system (1) is asymtotically equivalent to the given linear model (2).

More precisely, given $\alpha > 0$ and any (v_1,v_2,w,z), we construct (in a sense explicitely) a feedback u such that every possible state (x,y) for (1) corresponding to it (under any uncertain dynamics $g \in (G_1, G_2)'$) fulfils the model dynamics (2) up to an exponentially decaying error term dominated by (const.) exp ($-\alpha t$), $t \geq T$; moreover

$$|X(t) - W(t)| \leqslant (\text{const.}) \exp(-\alpha t), \quad t \geqslant T$$

where

$$X = (x, \dot{x}, y, \dot{y}) = (x_1, x_2, x_3, x_4), \quad W = (w, \dot{w}, z, \dot{z}) = (y_1, y_2, y_3, y_4),$$

and some T independent of X and explicitely estimated by known data.

Such a feedback u (in general discontinuous) is obtained by using variable structure control methods (see [1] recently extended to non linear control systems by the authors (see [2] and [3]), provided a set of explicit inequalities is satisfied by the known bounds $g_i^{\pm}$, $i = 1,2$, as follows. Let c_1, c_2 such that $\alpha \leqslant \min(c_1,c_2)$.

Then put $e = X - W$ and

(3) $s_1(e) = c_1 e_1 + e_2$, $s_2(e) = c_2 e_3 + e_4$.

Consider now

$$p(x, y, v) = c_1 (y_2 - x_2) + a_1 y_2 + a_2 y_1 + a_3 v_1,$$

$$q(x, y, v) = c_2 (y_4 - x_4) + b_1 y_4 + b_2 y_3 + b_3 v_2.$$

Then we assume existence of some $u \in U$ fulfilling

$$(4) \begin{cases} g_1^- \geqslant p + k^2 \quad \text{if } s_1 > 0; \quad g_1^+ \leqslant p - k^2 \quad \text{if } s_1 < 0; \\ g_2^- \geqslant q + k^2 \quad \text{if } s_2 > 0; \quad g_2^+ \leqslant q - k^2 \quad \text{if } s_2 < 0. \end{cases}$$

These results may be generalized to higher-order control systems of the following form

$$\ddot{x}_i \in G_i (t, x_1, \dot{x}_1, \ldots, x_n, \dot{x}_n, u_1, \ldots, u_n), \quad i = 1,\ldots,n.$$

It is then likely that the number of the required inequalities corresponding to (4) may be reduced by using results of [7] .

The chattering effects due to the discontinuous nature of the asymptotically linearizing feedback may be reduced by appropriately combining results from [8], [9] , [10] . Moreover sufficient conditions may be obtained assuring that the feedback u is piecewise continuous (not only measurable).

An important property of the variable structure control systems (1), (3) is approximability (see [2] for the precise definition).

In essence this means that whenever some error vectors e_ε depend on disturbances described by the real parameters ε and satisfy

$$s_i(e_\varepsilon) \longrightarrow 0 \text{ as } \varepsilon \longrightarrow 0, \quad i = 1,2,$$

uniformly on compact intervals of $[T, +\infty)$, then $e_\varepsilon \longrightarrow e_o$ in the same sense, where e_o is uniquely defined and fulfils the sliding condition $s_i(e_o) = 0, i = 1,2$. Thus approximability prevents ambiguous behaviour in the sliding mode. Sufficient conditions can be found about the available data in order to fulfil such a property.

A PARTICULAR CASE. Let the uncertain control system be described by a single differential inclusion of order n with scalar control u, given by

$$\dot{x}_1 = x_2 , \dot{x}_2 = x_3, \ldots, \dot{x}_{n-1} = x_n, \dot{x}_n \in G(t, x, u) , u \in U.$$

Suppose that the model is given by

$$\dot{y}_1 = y_2 , \dot{y}_2 = y_3, \ldots, \dot{y}_{n-1} = y_n, \dot{y}_n = \sum_{i=1}^{n} a_i y_i + by ,$$

and let

$$G = [g^-, g^+] .$$

Fix real number $c_1,\ldots,c_{n-1}$ such that the polynomial (in h)

$$h^{n-1} + c_{n-1} h^{n-2} + \ldots + c_2 h + c_1$$

is Hurwitz. Denote by

$$e = y - x$$

the error vector and set

$$s(e) = e_n + \sum_{i=1}^{n-1} c_i e_i ,$$

$$p(x, y, v) = bv + \sum_{i=1}^{n} a_i y_i + \sum_{i=1}^{n-1} c_i (y_{i+1} - x_{i+1}).$$

The asymptotic linear behaviour (in the sense defined above) obtains provided the following holds. For any model control-state pair y,v, every $t \geq 0$ and every x we can find $u \in U$ such that

$$g^-(t, x, u) \geq k^2 + p(x, y, v) \text{ if } s(y - x) > 0,$$
$$g^+(t, x, u) \leq - k^2 + p(x, y, v) \text{ if } s(y - x) < 0,$$

for some fixed constant $k \neq 0$.

Arbitrary exponential decay of the error term is obtained by a proper choice of $c_1, \ldots,c_{n-1}$. See [4] , [5] for more details, examples and comparisons about this particular case.

REFERENCES

[1] V.I. Utkin: "Sliding modes and their application in variable structure systems", MIR, Moscow, 1978.

[2] G.Bartolini, T.Zolezzi: "Control of non linear variable structure systems", J. Math. Anal. Appl. 118 (1986), 42-62.

[3] G.Bartolini, T.Zolezzi: "Variable structure systems non linear in the control law", IEEE Trans. Autom. Control 30 (1985), 681-684.

[4] G.Bartolini, T.Zolezzi: "Asymptotic linearization by variable structure control", Proc. 25th IEEE CDC (Athens, 1986), 2061-2062.

[5] G.Bartolini, T.Zolezzi: "Asymptotic linearization of uncertain systems by variable structure control", Systems Control Lett. 10 (1988), 111-117.

[6] B.E.Paden, S.S.Sastry: "A calculus for computing Filippov's differential inclusion with application to the variable structure control of robot manipulators", IEEE Trans. Circuits systems 34 (1987), 73-82.

[7] S.Baida, D.Izosimov: "Vector methods of design of sliding motion and simplex algorithms", Autom. Telemekh. 7 (1985), 56-63.

[8] J.A.Burton, A.S.Zinober: "Continuous approximation of variable structure control", Int. J. Systems Sci. 17 (1986), 875-885.

[9] A.Balestrino, G. De Maria, L.Sciavicco; J.Dyn. Systems Meas. Control 105 (1983), 141-151..

[10] G.Bartolini:"A note on *chattering* phenomenon in discontinuous control systems", (in preparation).

OUTPUT FEEDBACK CONTROL OF UNCERTAIN SYSTEMS IN THE PRESENCE OF UNMODELED ACTUATOR AND SENSOR DYNAMICS

Stanislaw H. Żak and Mehrez Hached

School of Electrical Engineering
Purdue University
West Lafayette, IN 47907
USA

ABSTRACT

This paper analyzes the performance of output feedback controllers for a class of uncertain time-varying nonlinear systems in the presence of unmodeled actuator and sensor dynamics. In particular, on the basis of known nominal model and bounds on the uncertainties, and initially neglecting actuator and sensor dynamics, high-gain output feedback schemes are determined which force the output to track a given signal. Then, the effects of actuator and sensor dynamics are investigated on the performance of the tracking system.

KEY WORDS: Nonlinear systems, Output feedback, Uncertain systems, Singular perturbations.

1. INTRODUCTION

Recently, major progress has been made in the analysis and design of nonlinear control systems. Different approaches have been proposed (Utkin, [1], [2], Corless and Leitmann [17], Hunt et al. [5], Su et al. [6], Glad [22], [23], Bauman and Rugh [19], DeCarlo et al. [10], Isidori [15], Walcott and Żak [8], Steinberg and Corless [12]). An important property of control systems is their robustness, i.e. the ability of the system to retain certain performance measures in the presence of perturbations. Or in other words; "the ability of a control system to function even when the actual system differs from the model used for designing the controller" (Glad [22]). The system model used by the designer may differ from the controlled system because of model uncertainties or neglected high-frequency dynamics. Specifically, when devising a model of the plant, small time constants corresponding to actuator and/or sensor dynamics are neglected. Furthermore, it is often impossible to measure directly all the components of the state or output vectors. In order to restore them additional sensors are used which lead to motions different from the motions predicted by the plant model.

The problem of controlling a system in the presence of unmodeled actuator and sensor dynamics has received recently the attention of many researchers. In particular Bondarev et al. [7], and Żak et al. [25] studied the influence of neglected high-frequency dynamics on the variable structure control systems. Leitmann et al. [9] studied the robustness with respect to neglected actuator and sensor dynamics of state feedback controllers for uncertain systems. Glad [23] considered the sensitivity of the system to variations in gain at the input, corresponding to nonideal behavior of the actuators. The problem of the robustness of various output feedback control algorithms based on a reduced-order model with neglected high-frequency dynamics was investigated by O'Reilly [18] and Vostrikov et al. [24] using singular perturbation techniques.

The purpose of this paper is to analyze the effect of neglected high-frequency dynamics on various output feedback control designs for nonlinear uncertain systems. Our approach is inspired by Marino [4], Utkin [2], and Vostrikov et al. [24]. The tools we use in this paper are the high-gain output feedback and Lie derivatives.

The paper is organized as follows. Section 2 is devoted to the description of the class of nonlinear systems we consider along with the problem statement. The next section presents some background material and preliminary results. The following sections discuss different high-gain output feedback control schemes. Then the effects on the performance of the closed-loop system of unmodeled actuator and sensor dynamics are investigated. Finally, Section 6 contains concluding remarks.

2. PROBLEM STATEMENT

In this paper we consider a class of dynamical systems governed by the following equations

$$\left.\begin{aligned}\dot{x}(t) &= f(t,x) + G(t,x)\,[u(t) + \xi(t,x)]\\ y(t) &= h(x)\,,\end{aligned}\right\} \qquad (2.1)$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^m$, and $\xi(\cdot)\ \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the lumped uncertain element. We assume that the norm of the uncertain element is bounded by a known bounded nonnegative function; that is for all $(t,x) \in \mathbb{R} \times \mathbb{R}^n$

$$\|\xi(t,x)\| \leq \rho(t,x)\,,$$

where $\rho(\cdot) : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}_+$, and $\|\cdot\|$ is the Euclidean norm i.e., $\|x\| = (\sum_{i=1}^{n} |x_i|^2)^{1/2}$.

Note that the only information assumed about the uncertain vector is its maximum possible energy. If the uncertainties $\xi(t,x)$ enter structurally into the state equations as in (2.1) then we say that the matching condition is satisfied [17].

The function $f(\cdot)$ is a continuous single-valued vector-function and $G(\cdot)$ is a continuous single-valued matrix function with rank $G = m$. Furthermore, we require that $f(t,0) = 0$ for all t. The output vector function $h(\cdot)$ is continuously differentiable and

$h(0) = 0$.

In this paper we analyze two different output feedback control strategies. The first is the high-gain output feedback stabilization scheme. In the synthesis of this control law we utilize a nonlinear transformation which brings the original system into the "regular form" ([20]) from where the design is performed.

The aim of the second control law is to ensure the tracking property of the output of some given reference signal.

For both control strategies we will investigate the effects of the unmodeled actuator and sensor dynamics on the performance of the closed-loop systems.

3. PRELIMINARY RESULTS

LIE DERIVATIVES

Time-Invariant Lie Derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ be C^∞ vector fields on $\mathbb{R}^n$. The Lie bracket is defined by

$$[f,g] \triangleq \frac{\partial f}{\partial x} g - \frac{\partial g}{\partial x} f ,$$

where $\frac{\partial f}{\partial x}$ and $\frac{\partial g}{\partial x}$ are the Jacobian matrices of f and g, respectively. Using an alternative notation, one can represent the Lie bracket as follows

$$[f,g] = (ad^1 f,g) .$$

Also, define

$$(ad^k f,g) = [f,(ad^{k-1} f,g)] ,$$

where, by definition

$$(ad^0 f,g) = g .$$

Next, consider a C^∞ function $h : \mathbb{R}^n \to \mathbb{R}$. Let $dh = \nabla^T h$ be the derivative of h with respect to x, where ∇h is the gradient of h with respect to x. Then the Lie derivative of h with respect to f is defined by

$$L_f h = L_f(h) = <dh,f> = \nabla^T h \cdot f .$$

The following notation is employed throughout this paper

$$
\begin{aligned}
L_f^0 h &= h \\
L_f^1 h &= L_f h \\
&\vdots \\
L_f^k h &= L_f(L_f^{k-1} h) .
\end{aligned}
$$

The Lie derivative of dh with respect to the vector field f is defined by

$$L_f(dh) = \left[\frac{\partial (dh)^T}{\partial x} f\right]^T + (dh)\,\frac{\partial f}{\partial x} .$$

One may easily verify that these three Lie derivatives obey the following so-called Leibnitz formula

$$L_{[f,g]} h = <dh, [f,g]> = L_g L_f h - L_f L_g h .$$

Furthermore, the following relation is valid

$$dL_f h = L_f(dh) .$$

Time-Varying Lie Derivatives

Suppose now f and g are C^∞ time-varying vector fields, i.e. $f(\cdot) : \mathbb{R}\times\mathbb{R}^n \rightarrow \mathbb{R}^n$, $g(\cdot) : \mathbb{R}\times\mathbb{R}^n \rightarrow \mathbb{R}^n$. Then the time-varying Lie bracket is defined by

$$(\Gamma^1 f, g) \triangleq (ad^1 f, g) - \frac{\partial g}{\partial t} ,$$

and

$$(\Gamma^k f, g) = (\Gamma^1 f, (\Gamma^{k-1} f, g)) ,$$

where

$$(\Gamma^0 f, g) \triangleq g .$$

Next consider a C^∞ function $h(\cdot) : \mathbb{R}\times\mathbb{R}^n \rightarrow \mathbb{R}$. Then the time-varying Lie derivative of h with respect to f is defined by

$$\mathscr{L}_f h = \mathscr{L}_f(h) \triangleq L_f h + \frac{\partial h}{\partial t} .$$

We define

$$
\begin{aligned}
&\mathscr{L}_f^0 h \triangleq h , \\
&\mathscr{L}_f^k h \triangleq \mathscr{L}_f(\mathscr{L}_f^{k-1} h) = L_f(\mathscr{L}_f^{k-1} h) + \frac{\partial \mathscr{L}_f^{k-1} h}{\partial t} .
\end{aligned}
$$

The time-varying Lie derivative of dh with respect to the time-varying vector field f is defined by

$$\mathscr{L}_f dh = f^T \left(\frac{\partial(dh)}{\partial x}\right) + (dh)\frac{\partial f}{\partial x} + \frac{\partial}{\partial t}(dh) .$$

Note that

$$d\mathscr{L}_f h = \mathscr{L}_f(dh) .$$

One may verify that the above defined time-varying Lie derivatives obey the following formula

$$\begin{aligned} &< dh, (\Gamma^1 f, g) > \\ &= L_{(\Gamma^1 f, g)} h = L_g L_f h - L_f L_g h - L_{\frac{\partial g}{\partial t}} h \\ &= L_g \mathscr{L}_f h - \mathscr{L}_f L_g h . \end{aligned}$$

MARKOV PARAMETERS

The affine Markov parameters are defined as the elements of the matrix resulting from the product of the observability and controllability matrices of an affine nonlinear system described by the following equations

$$\left.\begin{aligned} \dot{x} &= f(t,x) + g_1(t,x)u_1 + \dots + g_m(t,x)u_m \\ y &= h(x) = [h_1(x) , \dots , h_p(x)]^T , \end{aligned}\right\} \tag{3.1}$$

where $f, g_1, \dots, g_m$: $\mathbb{R}\times\mathbb{R}^n \longrightarrow \mathbb{R}^n$ and h : $\mathbb{R}^n \longrightarrow \mathbb{R}^p$ are C^∞ vector fields.

The observability matrix of such a system is defined by the following $(np)\times n$ matrix

$$\mathcal{O} = \begin{bmatrix} dh_1 \\ \vdots \\ dh_p \\ \mathcal{L}_f(dh_1) \\ \vdots \\ \mathcal{L}_f(dh_p) \\ \vdots \\ \mathcal{L}_f^{n-1}(dh_1) \\ \vdots \\ \mathcal{L}_f^{n-1}(dh_p) \end{bmatrix} . \tag{3.2}$$

The controllability matrix is defined by the following n×(nm) matrix

$$\mathcal{C} = \left[g_1,\dots,g_m,(\Gamma^1 f,g_1),\dots,(\Gamma^1 f,g_m),\dots,(\Gamma^{n-1} f,g_1),\dots,(\Gamma^{n-1} f,g_m) \right] . \tag{3.3}$$

So the elements of the matrix $\mathcal{O}\mathcal{C}$ have the form

$$\begin{aligned} (\mathcal{L}_f^j(dh_\jmath))\,(\Gamma^i f,g_\alpha) &= <\mathcal{L}_f^j(dh_\jmath),\,(\Gamma^i f,g_\alpha)> \\ &= <d\mathcal{L}_f^j h_\jmath\;,\,(\Gamma^i f,g_\alpha)> \\ &= L_{(\Gamma^i f,g_\alpha)}\,\mathcal{L}_f^j h_\jmath \end{aligned} \tag{3.4}$$

for i,j = 1,...,n−1, α = 1,...,m , $\jmath$ = 1,...,p, and are referred to as the affine Markov parameters.

Theorem 3.1: If there exist constants c_k, k = 0,1,... such that the Markov parameters satisfy

$$L_{(\Gamma^i f,g_\alpha)}\,\mathcal{L}_f^j h_\jmath = c_k = c_{i+j} \;, \tag{3.5}$$

then

$$L_{(\Gamma^i f,g_\alpha)}\mathcal{L}_f^j h_\jmath = L_{g_\alpha}\mathcal{L}_f^{i+j} h_\jmath = \text{const} = c_{i+j} \;.$$

Proof: Repeated application of the definitions of Lie derivatives and condition (3.5) yields the following

$$\begin{aligned} L_{(\Gamma^i f,g_\alpha)}\mathcal{L}_f^j h_\jmath &= <d\mathcal{L}_f^j h_\jmath\;,\,(\Gamma f,(\Gamma^{i-1} f,g_\alpha))> \\ &= L_{(\Gamma^{i-1} f,g_\alpha)}\mathcal{L}_f\mathcal{L}_f^j h_\jmath - \mathcal{L}_f L_{(\Gamma^{i-1} f,g_\alpha)}\mathcal{L}_f^j h_\jmath \end{aligned}$$

$$= L_{(\Gamma^{i-1}f, g_o)} \mathscr{L}_f^{j+1} h_\beta - \mathscr{L}_f c_{i+j-1}$$

$$= L_{(\Gamma^{i-1}f, g_o)} \mathscr{L}_f^{j+1} h_\beta \ .$$

Continuing in this manner we find that

$$L_{(\Gamma^{i}f, g_o)} \mathscr{L}_f^{j} h_\beta = L_{g_o} \mathscr{L}_f^{i+j} h_\beta = \text{const} = c_{i+j} \ .$$

□

For further information about Markov parameters for nonlinear time-invariant systems the reader is referred to [11] and [14].

Consider now a plant modeled by (3.1), where p = m, and the high gain control law

$$u = k\, s(x) \tag{3.6}$$

where $k > 0$ is a scalar and the function $s(\cdot) : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is continuously differentiable. Assume that $\det SG \neq 0$, where

$$S = \frac{\partial s}{\partial x} \ , \ \text{ and } \ G = [g_1, g_2, \ldots, g_m] \ .$$

Then we have

Theorem 3.2 ([2],[21]):

If

(i) the functions $f(t,x)$, $G(t,x)s(x)$, and $f_0 = f - G(SG)^{-1}Sf$ satisfy Lipschitz conditions for all x

(ii) the system

$$\frac{ds}{dt} = (SG)s$$

is uniformly exponentially stable, that is there exist positive $A \geqq 1$ and α such that

$$||s(x)|| < A||s(x(0))||e^{-\alpha t} \ ,$$

then for any positive Δ, and T there exists a positive k_0 such that

$$||s(x(t))|| < \Delta$$

for $k > k_0$ and $t_0 + t_1 < t < T$ on the solutions of (3.1) with the control $u = ks(x)$, and $\lim_{k \to \infty} t_1 = 0$.

4. THE OUTPUT REGULATION PROBLEM

Consider the nominal system, that is the system without uncertainty as described by

$$\left.\begin{aligned} \dot{x} &= f(t,x) + G(t,x)u \\ y &= h(x) . \end{aligned}\right\} \tag{4.1}$$

First we define the decoupling indices for the system (4.1). We consider each of the m output channels separately. So considering the first output channel we form the following row vector which we will call the decoupling vector for channel one

$$[L_{g_1}h_1,\ L_{g_2}h_1,\dots,L_{g_m}h_1] . \tag{4.2}$$

If this row vector is not identically equal to zero, then we define the decoupling index of the first channel to be zero, or $d_1 = 0$.

However, if the row vector is identically equal to zero we proceed to form the following decoupling vector

$$[L_{g_1}\mathscr{L}_f h_1,\ L_{g_2}\mathscr{L}_f h_1\ ,\ \dots\ ,\ L_{g_m}\mathscr{L}_f h_1] .$$

Again we determine if it is identically equal to zero, or not. If it is not we stop and define $d_1 = 1$. If it is zero we proceed further by forming

$$[L_{g_1}\mathscr{L}_f^2 h_1,\ L_{g_2}\mathscr{L}_f^2 h_1\ ,\ \dots\ ,\ L_{g_m}\mathscr{L}_f^2 h_1] ,$$

and so on.

So the decoupling index of channel 1, is equal to the smallest integer d_1 for which the decoupling vector,

$$[L_{g_1}\mathscr{L}_f^{d_1} h_1,\ L_{g_2}\mathscr{L}_f^{d_1} h_1\ ,\ \dots\ ,\ L_{g_m}\mathscr{L}_f^{d_1} h_1] ,$$

is not identically equal to zero.

Similar procedure for the other output channels yields a set of m parameters, d_i for $i = 1,2,\dots m$.

The decoupling indices are an indication of what the lowest derivative of each output channel needs to be utilized for an output control to be effective. By taking the time derivative of the i$^{\text{th}}$ output channel we obtain

$$\dot{y}_i = \frac{\partial h_i}{\partial x}\dot{x} = \frac{\partial h_i}{\partial x}\,(f + g_1 u_1 + \dots + g_m u_m) ,$$

hence

$$\dot{y}_i = L_f h_i + [L_{g_1}h_i\ ,\ \dots\ ,\ L_{g_m}h_i]\ u .$$

Thus if $[L_{g_1}h_i,\ L_{g_2}h_i\ ,\ \dots\ ,\ L_{g_m}h_i] = [0]$ then u has no effect on the output y_i, so we need to form $\ddot{y}_i$ where

$$\ddot{y}_i = \frac{\partial \dot{y}_i}{\partial x}\dot{x} = \mathscr{L}_f^2 h_i + [L_{g_1}\mathscr{L}_f h_i \ , \ \dots \ , \ L_{g_m}\mathscr{L}_f h_i]u \ .$$

Again if $[L_{g_1}\mathscr{L}_f h_i, L_{g_2}\mathscr{L}_f h_i \ , \ \dots \ , \ L_{g_m}\mathscr{L}_f h_i] = [0]$, then u has no effect on the output and we need to take higher derivatives of y_i in a similar fashion as before.

Now that we have obtained the set of decoupling indices, we consider all the output channels together to form the following matrix

$$N = \begin{bmatrix} L_{g_1}\mathscr{L}_f^{d_1}h_1 & \cdots & L_{g_m}\mathscr{L}_f^{d_1}h_1 \\ \vdots & & \vdots \\ L_{g_1}\mathscr{L}_f^{d_m}h_m & \cdots & L_{g_m}\mathscr{L}_f^{d_m}h_m \end{bmatrix} . \tag{4.3}$$

We will assume that the matrix N is nonsingular and we will further assume that the Markov parameters of the system (4.1) are constant. Hence by the virtue of Theorem 3.1 the matrix N is constant.

With the N matrix constant and nonsingular, we proceed to construct a high-gain output control which will regulate the output to zero.

We will consider two cases. The first case is when all decoupling indices are equal to zero, and the second case when some, or all, decoupling indices are not equal to zero.

For a rigorous treatment of the decoupling problem for nonlinear time-invariant systems the reader is referred to [3], [14], [16].

Case 1: For this case the N matrix will have the following form

$$N = \begin{bmatrix} L_{g_1}h_1 & \cdots & L_{g_m}h_1 \\ \vdots & & \vdots \\ L_{g_1}h_m & \cdots & L_{g_m}h_m \end{bmatrix} = \frac{\partial h}{\partial x}G = HG \ , \tag{4.4}$$

where $\frac{\partial h}{\partial x}$ is the Jacobian matrix of h and $G = [g_1, \dots, g_m]$.

If we employ the following diffeomorphic state variable transformation

$$\left.\begin{aligned} \bar{x}_1 &= \phi_1(t,x) \\ \bar{x}_2 &= \phi_2(t,x) \\ &\vdots \\ \bar{x}_{n-m+1} &= h_1(x) \\ &\vdots \\ \bar{x}_n &= h_m(x)\,, \end{aligned}\right\} \tag{4.5}$$

where the ϕ_i's are chosen such that

$$L_{g_j}\phi_i = 0\,,\ \ j = 1,\ldots,m \ \text{ for all } \ i = 1,\ldots,n-m, \tag{4.6}$$

then the system (2.1) in the new coordinates will have the following form

$$\begin{bmatrix} \dot{\bar{x}}_1 \\ \vdots \\ \dot{\bar{x}}_{n-m} \end{bmatrix} = \begin{bmatrix} \mathcal{L}_f\phi_1 \\ \vdots \\ \mathcal{L}_f\phi_{n-m} \end{bmatrix} \tag{4.7a}$$

$$\begin{bmatrix} \dot{\bar{x}}_{n-m+1} \\ \vdots \\ \dot{\bar{x}}_n \end{bmatrix} = \begin{bmatrix} L_f h_1 \\ \vdots \\ L_f h_m \end{bmatrix} + \begin{bmatrix} L_{g_1}h_1 & \cdots & L_{g_m}h_1 \\ \vdots & & \vdots \\ L_{g_1}h_m & \cdots & L_{g_m}h_m \end{bmatrix}(u + \xi)\,. \tag{4.7b}$$

We will now employ the high gain output feedback control as given by

$$u = \frac{1}{\epsilon}\,K^{\circ}h(x)\,, \tag{4.8}$$

where ϵ is a small constant and K° is an $m{\times}m$ constant matrix. Under the influence of this control, the system equations become

$$\begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_{n-m} \end{bmatrix} = \begin{bmatrix} \mathcal{L}_f\phi_1 \\ \vdots \\ \mathcal{L}_f\phi_{n-m} \end{bmatrix} \tag{4.9a}$$

$$\begin{bmatrix} \dot{\bar{x}}_{n-m+1} \\ \vdots \\ \dot{\bar{x}}_n \end{bmatrix} = \begin{bmatrix} L_f h_1 \\ \vdots \\ L_f h_m \end{bmatrix} + \frac{1}{\epsilon} N K^{o} h(x) + N\xi . \tag{4.9b}$$

We see that the application of this control decouples the system into the slow and fast subsystems. The dynamics of the slow subsystem are given by

$$\left. \begin{aligned} \begin{bmatrix} \dot{\bar{x}}_1 \\ \vdots \\ \dot{\bar{x}}_{n-m} \end{bmatrix} &= \begin{bmatrix} \mathscr{L}_f \phi_1 \\ \vdots \\ \mathscr{L}_f \phi_{n-m} \end{bmatrix} \\ y = h(x) &= \begin{bmatrix} \bar{x}_{n-m+1} \\ \vdots \\ \bar{x}_n \end{bmatrix} = 0 , \end{aligned} \right\} \tag{4.10}$$

whereas by invoking the following change in the time variable,

$$t = \epsilon \tau , \tag{4.11}$$

the equations describing the dynamics of the fast system are given by

$$\frac{d}{d\tau} \begin{bmatrix} \bar{x}_{n-m+1} \\ \vdots \\ \bar{x}_n \end{bmatrix} = \epsilon \begin{bmatrix} \mathscr{L}_f h_1 \\ \vdots \\ \mathscr{L}_f h_m \end{bmatrix} + N K^{o} h(x) + \epsilon N \xi , \tag{4.12}$$

and for sufficiently small ϵ, the above equations simplify to

$$\frac{d}{d\tau} \begin{bmatrix} \bar{x}_{n-m+1} \\ \vdots \\ \bar{x}_n \end{bmatrix} = N K^{o} h(x) = N K^{o} \begin{bmatrix} \bar{x}_{n-m+1} \\ \vdots \\ \bar{x}_n \end{bmatrix} . \tag{4.13}$$

Observing that the part of our transformation in (4.5) is $y = [\bar{x}_{n-m-1} , \dots , \bar{x}_n]^T$, we can rewrite the above equation as

$$\frac{dy}{d\tau} = N K^{o} y . \tag{4.14}$$

Note that by an appropriate choice of the matrix K^{o} the fast subsystem can be made uniformly exponentially stable. Thus if J is the required uniformly exponentially stable matrix then,

$$K^{o} = N^{-1}J\,, \tag{4.15}$$

and K^{o} can be evaluated since N is assumed to be nonsingular.

By invoking Theorem 3.2 we see that the stability of the fast subsystem will result in the trajectories of the system (4.7) converging to the Δ-vicinity of the manifold $y(x) = 0$. Thus the output is regulated to zero. Within the Δ-vicinity of the manifold, the system will be governed by equation (4.10) which represents the dynamics of the slow subsystem. From equation (4.10), we notice that we do not have any influence on the internal stability of the slow subsystem when the output is regulated to zero. We assume however that the slow subsystem is asymptotically stable. The stability of the slow subsystem is a structural property of the plant. This subject requires further research.

Although Theorem 3.2 was stated for nonlinear systems without uncertainties, it also applies to our particular case. This is because the uncertainties in the system (2.1) are bounded by a known bounded function.

Case 2: Let us first reorder the output channels so that they are ordered in ascending values of their decoupling indices. Thus y_1 is assigned to the channel with the smallest d_i, and y_m to the one with the largest d_i.

We then employ the following diffeomorphic state variable transformation

$$\bar{x} = \begin{bmatrix} \phi_1(t,x) \\ \phi_2(t,x) \\ \vdots \\ h_m(x) \\ \vdots \\ \mathscr{L}_f^{d_1-1}h_1 \\ \vdots \\ \mathscr{L}_f^{d_m-1}h_m \\ \mathscr{L}_f^{d_1}h_1 \\ \vdots \\ \mathscr{L}_f^{d_m}h_m \end{bmatrix}, \tag{4.16}$$

where the ϕ_i's are chosen such that

$$L_{g_j}\phi_i = 0 \quad j = 1,\ldots,m\,.$$

The system (2.1) in the new coordinates will have the following form

$$\left.\begin{aligned}\dot{\bar{x}}^1 &= f_1(t,\bar{x})\\ \dot{\bar{x}}^2 &= f_2(t,\bar{x}) + N(u+\xi)\,,\end{aligned}\right\} \qquad (4.17)$$

where $\bar{x}^1 \in \mathbb{R}^{n-m}$, $\bar{x}^2 \in \mathbb{R}^m$, and N is given by (4.3). The existence conditions of the transformation (4.16) can be deduced from the results of [5], [20], [26], [27].

Note that

$$\dot{\bar{x}}_i^2 = y_i^{(d_i)} \quad , \quad i = 1,\ldots,m\,,$$

where $(\ )^{(j)}$ denotes the j-th derivative of ( ) with respect to t. The control law will have the form

$$u = \frac{1}{\epsilon} N^{-1} \begin{bmatrix} k_{1,1}y_1 + \ldots + k_{1,d_1+1}y_1^{(d_1)} \\ \vdots \\ k_{m,1}y_m + \ldots + k_{m,d_m+1}y_m^{(d_m)} \end{bmatrix}. \qquad (4.18)$$

In the new coordinates the closed-loop system (4.17), (4.18) is decoupled into the slow and fast subsystems. The slow subsystem is governed by the equations

$$\left.\begin{aligned}\dot{\bar{x}}^1 &= f_1(t,\bar{x})\\ y &= 0\,.\end{aligned}\right\} \qquad (4.19)$$

As in the previous case, we have no influence on the stability of the slow subsystem. Therefore for the controller to be effective we have to assume that the system (2.1) without uncertainties is asymptotically stable when restricted to the manifold $y = 0$ which is equivalent to requirement that the system (4.19) is asymptotically stable.

As with regard to the fast subsystem we utilize a change in the time variable $t = \epsilon\,\tau$ to obtain

$$\begin{bmatrix} y_1^{(d_1+1)} \\ \vdots \\ y_m^{(d_m+1)} \end{bmatrix} = \epsilon(f_2 + N\xi) + \begin{bmatrix} k_{1,1}y_1 + \ldots + k_{1,d_1+1}y_1^{(d_1)} \\ \vdots \\ k_{m,1}y_m + \ldots + k_{m,d_m+1}y_m^{(d_m)} \end{bmatrix}. \qquad (4.20)$$

If we now choose k_{ij} in such a way that the simplified fast subsystem

$$\begin{bmatrix} y_1^{(d_1+1)} \\ \vdots \\ y_m^{(d_m+1)} \end{bmatrix} = \begin{bmatrix} k_{1,1}y_1 + \dots + k_{1,d_1+1}y_1^{(d_1)} \\ \vdots \\ k_{m,1y_m} + \dots + k_{m,d_m+1}y_m^{(d_m)} \end{bmatrix}$$

is uniformly exponentially stable then by the virtue of Theorem (3.1) the closed-loop system is asymptotically stable.

The above output feedback stabilization schemes are quite restrictive. Their effectiveness depends on the stability of the nominal system ($\dot{x} = f + Gu$) when restricted to the manifold $y = h(x) = 0$. In the following section we provide a more effective control scheme. Before that however, we will analyze the effect of unmodeled actuator dynamics on the performance of the closed-loop system with the high-gain output feedbacks.

SYSTEMS WITH FAST UNMODELED MOTIONS

We now investigate the effects of the introduction of uncertain actuator dynamics on the performance of the system (2.1) with high gain output feedback controllers.

Case 1: We will assume that the actuator dynamics is modeled by the following equation

$$\mu_a \dot{r} = Lr + M\bar{u}\ , \quad u = Nr + R\bar{u}\ , \ \bar{u} = c\,K^o y\ , \tag{4.21}$$

where $r \in \mathbb{R}^q$, $q \geqq m$, L is a Hurwitz matrix, μ_a is a positive constant that reflects the "fastness" of the actuator, the matrices L, M, R, and N satisfy the condition $R - NL^{-1}M = I_m$, and $c = \frac{1}{\epsilon}$ is a large constant.

Proposition 4.1: If the matrix L is Hurwitz, (the fast subsystem described by (4.21) is exponentially stable) then as μ_a approaches 0, the motion of the slow subsystem is described by (2.1) with $u = \bar{u} = c\,K^o y$.

Proof: The fast subsystem is described by (4.21). Replacing $\bar{u}$ by its value yields

$$\mu_a \dot{r} = Lr + cMK^o y\ . \tag{4.22}$$

Let $\tau = \mu_a^{-1} t$, hence (4.22) becomes

$$\frac{dr}{d\tau} = Lr + cMK^o y\ . \tag{4.23}$$

Since L is a Hurwitz matrix, then as τ approaches infinity we have $y =$ constant and

$$\lim_{r\to\infty} r = - cL^{-1}MK^{\circ}y ,$$

hence

$$u = Nr + cRK^{\circ}y = [- NL^{-1}M + R]cK^{\circ}y = cK^{\circ}y . \tag{4.24}$$

The expression for u as per (4.24) can also be found by setting $\mu_a = 0$. Hence the slow subsystem is described by (2.1) and (4.24).

□

Case 2: If the actuator dynamics for this case is also described by (4.21) with

$$\bar{u} = cN^{-1} \begin{bmatrix} K_{1,1y_1} + \dots + K_{1,d_1+1}y_1^{(d_1)} \\ \vdots \\ K_{m_1,y_m} + \dots + K_{m,d_m+1}y_m^{(d_m)} \end{bmatrix} ,$$

then using a similar argument as in the previous case we conclude that the slow subsystem is described by (2.1) with $u = \bar{u}$.

In conclusion, for a sufficiently fast actuator the proposed control schemes will stabilize the output.

5. THE TRACKING PROBLEM

Our goal now is to design a controller such that the output of the system (2.1) will track a given reference signal.

A sufficient condition for the output y to track the reference signal $\nu(t)$ is

$$\frac{d}{dt}[y - \nu(t)] = V[y - \nu(t)] \triangleq F(y, \nu(t)) , \tag{5.1}$$

where V is a Hurwitz matrix. If $\nu(t) =$ constant, then (5.1) becomes $\dot{y} = V[y - \nu]$.

We require that the closed-loop system (2.1) be asymptotically stable with respect to the time-varying manifold

$$\Omega = \{x : h(x(t)) - \nu(t) = y(t) - \nu(t) = 0\} .$$

The projection of the overall system on this manifold is

$$\begin{aligned} \dot{y}(t) - \dot{\nu}(t) &= H\dot{x} - \dot{\nu}(t) \\ &= Hf + HG(u + \xi) - \dot{\nu}(t) . \end{aligned}$$

Using equation (5.1) and solving for u, we obtain the following control law

$$\bar{u} = (HG)^{-1}[F(y, \nu(t)) - Hf + \dot{\nu}(t)] - \xi . \tag{5.2}$$

In order to implement the control law (5.2), we would have to have the exact knowledge of the uncertain vector $\xi(t,x)$. Hence this control strategy is impractical. In what follows we propose a practical control algorithm which approximates the controller (5.2).

Consider the following control strategy

$$u = K[V(y - \nu(t)) - (\dot{y} - \dot{\nu}(t))] , \tag{5.3}$$

where K is the matrix of gain coefficients, $K = cK^{\circ}$, and c is a scalar large factor. At the present time, we will assume that $\dot{y}$ can be measured exactly. Later, we will investigate the case in which $\dot{y}$ is measured by a sensor.

To analyze the behavior of the system (2.1) with the control law (5.3) in the presence of unmodeled actuator and sensor dynamics we will employ the arguments of Vostrikov et al. [24] used for systems without uncertainties.

Along the trajectories of the motion of the dynamical system (2.1), $\dot{y}$ is given by

$$\dot{y} = Hf(t,x) + HG(t,x)(u + \xi(t,x)) . \tag{5.4}$$

Proposition 5.1: If $\det(I + cHGK^{\circ}) \neq 0$, and $\det(HG) \neq 0$, then

(a) $\lim\limits_{c\to\infty} \dfrac{d}{dt} [y - \nu(t)] = F(y,\nu(t))$,

(b) $\lim\limits_{c\to\infty} u = (HG)^{-1} [F(y,\nu(t)) - Hf + \dot{\nu}(t) - HG\xi]$.

Proof: In what follows we shall utilize the arguments of Vostrikov et al. [24].

We first prove part (a). Recall that

$$\dot{y} = H\dot{x} \, , \; \dot{x} = f + G(K[F - (\dot{y} - \dot{\nu})] + \xi) ,$$

thus, we have

$$\dot{y} = Hf + HGK(F - \dot{y} + \dot{\nu}) + HG\xi ,$$

regrouping the $\dot{y}$ terms leads to

$$(I + HGK)\dot{y} = Hf + HGK(F + \dot{\nu}) + HG\xi ,$$

Hence, for $K = cK^{\circ}$, we have

$$\dot{y} = (I + cHGK^{\circ})^{-1}(Hf + HG\xi) + (I + cHGK^{\circ})^{-1}cHGK^{\circ}(F + \dot{\nu}) .$$

Taking $\lim\limits_{c\to\infty} \dot{y}$, the first term approaches zero, while the second term approaches $F + \dot{\nu}$, therefore

$$\lim_{c\to\infty} \dot{y} = F + \dot{\nu} .$$

We now prove part (b).

We have

$$u = K[F - (\dot{y} - \dot{\nu})] , \ \dot{y} = Hf + HG(u + \xi) ,$$

therefore

$$u = K[F - Hf - HG(u + \xi) + \dot{\nu}] .$$

Regrouping the u terms leads to

$$(I + KHG)u = K[F - Hf - HG\xi + \dot{\nu}] ,$$

hence, for $K = cK^{\circ}$, we have

$$u = (I + cK^{\circ}HG)^{-1} cK^{\circ} [F - Hf - HG\xi + \dot{\nu}] .$$

Thus

$$\lim_{c \to \infty} u = (HG)^{-1} [F - Hf - HG\xi + \dot{\nu}] .$$

□

SYSTEMS WITH FAST UNMODELED MOTIONS

We will now investigate the effects of the neglected actuator dynamics on the performance of the system (2.1) with the control law (5.3).

Suppose that the actuator dynamics is modeled by the following equation

$$\mu_a \dot{r} = Lr + M\bar{u} , \ u = Nr , \ \bar{u} = K(F - \dot{y} + \dot{\nu}) , \tag{5.5}$$

where $r \in \mathbb{R}^q$, $q \geqq m$, L is a Hurwitz matrix, μ_a is a positive constant that reflects the "fastness" of these dynamics, and the matrices L, M, N satisfy the condition $- NL^{-1}M = I$.

The system described by (2.1), and (5.5) may be studied by the methods of the theory of differential equations with small parameters in some of the derivatives [24].

For such systems, the overall motion can be decoupled into the fast and slow components [21] [24]. The method of decoupling motions is advantageous in systems involving high-gain feedback and/or singular perturbations. The main idea behind the theory is to decouple the system into two subsystems of lower dimensionality. The equations of the slow motions and the convergence conditions for the fast motions are examined in [21] and [24].

In the following proposition we investigate the effects of the actuator dynamics on the performance of the system (2.1).

Proposition 5.2: If the matrix (L − MKHGN) is a Hurwitz matrix, then as $c \to \infty$ the motion of the slow subsystem will be described by (2.1) with $u = \bar{u}$.

Proof: As $\mu_s \to 0$, the slow subsystem is described by the following equations

$$\dot{x} = f + G(u + \xi)\,, \quad u = \bar{u} = K(F - \dot{y} + \nu)\,.$$

We now examine the condition for the stability of the fast subsystem.

The fast subsystem is described by equation (5.5). Replacing $\bar{u}$ by its value yields

$$\mu_s \dot{r} = Lr + MK(F - Hf - HGNr - HG\xi + \nu)\,. \tag{5.6}$$

Let $\tau = \mu_s^{-1}t$, hence equation (5.6) becomes

$$\frac{dr}{d\tau} = (L - MKHGN)r + MK(F - Hf - HG\xi + \nu)\,. \tag{5.7}$$

where x = constant, t = constant.

If the matrix (L − MKHGN) is Hurwitz, then

$$\lim_{\tau \to \infty} r = -(L - MKHGN)^{-1}MK(F - Hf - HG\xi + \nu)\,.$$

Applying twice the following matrix identity know as the matrix inversion lemma

$$(A_{11} + A_{12}A_{22}A_{21})^{-1} = A_{11}^{-1} - A_{11}^{-1}A_{12}(A_{21}A_{11}^{-1}A_{12} + A_{22}^{-1})^{-1}A_{21}A_{11}^{-1}\,,$$

and the condition $-NL^{-1}M = I_m$ we obtain

$$\begin{aligned}
&\lim_{c \to \infty} N(L - MK^{\circ}(cHG)N)^{-1}cMK^{\circ} \\
&= \lim_{c \to \infty} cK^{\circ}[-I_m + (K^{\circ} + (cHG)^{-1})^{-1}K^{\circ}] \\
&= -(HG)^{-1}\,.
\end{aligned} \tag{5.8}$$

Hence

$$\lim_{c \to \infty} u = \lim_{c \to \infty} Nr = (HG)^{-1}[F - Hf + \nu] - \xi\,.$$

□

INFLUENCE OF SENSOR DYNAMICS

To implement the control law (5.3), the vector $\dot{y}$ has to be measured by a sensor (approximate differentiator). Suppose that the approximate differentiator is modeled by the following equation

$$\left.\begin{aligned} \mu_s \dot{z} &= Az + Dh(x) \\ \hat{y} &= Pz\,, \end{aligned}\right\} \tag{5.9}$$

where $z \in \mathbb{R}^q$, $\hat{y} \in \mathbb{R}^m$, $q \geq m$, and μ_s is a "small" parameter that reflects the "fastness" of the approximate differentiator, $\hat{y}$ is the estimate of y, A is a Hurwitz matrix, and the matrices P, A, and D satisfy the condition $-PA^{-1}D = I$. We shall also use $\dot{\hat{y}}$ instead of $\dot{y}$ in the control law (5.3). Therefore, we have

$$\dot{\hat{y}} = P\dot{z} = \mu_s^{-1}P(Az + Dh(x))\,.$$

Again, to examine the system (2.1) with the control strategy (5.3) and the approximate differentiator (5.9), we shall refer to the theory of decoupling motions ([21], [24]).

If we denote

$$s = Az + Dh(x) = \mu_s \dot{z}\,, \tag{5.10}$$

then the method of decoupling motions described in [21] is suitable for the resulting system. We now examine the condition for the convergence of the fast motions to the manifold s = 0.

The projection of the overall system on the manifold s is given by

$$\begin{aligned} \dot{s} &= A\dot{z} + D\dot{y} \\ &= A\dot{z} + DH\dot{x} \\ &= A\mu_s^{-1}s + DH[f + G(u + \xi)]\,. \end{aligned}$$

Replacing u by its value yields

$$\dot{s} = (A - DGHKP)\mu_s^{-1}s + DH(f + GKF + GK\dot{\nu} + G\xi)\,. \tag{5.11}$$

If we now multiply both sides of the above equation by μ_s, and let $t = \mu_s \tau$, we get

$$\frac{ds}{d\tau} = \mu_s \frac{ds}{dt} = (A - DHGKP)s + \mu_s DH(f + GKF + G\xi + GK\dot{\nu})\,,$$

where x = constant, t = constant. If the matrix [A − DHGKP] is Hurwitz then

$$\lim_{\tau \to \infty} s = -\mu_s[A - DHGKP]^{-1}[DHf + DHGK(F + \dot{\nu}) + DHG\xi]\,.$$

Using twice the matrix inversion lemma and the condition $-PA^{-1}D = I_m$ we obtain

$$\lim_{c \to \infty} (A - DHGKP)^{-1}$$

$$= \lim_{c \to \infty} P^{-1}P(A - D(cHG)K^{\circ}P)^{-1}$$

$$= \lim_{c \to \infty} P^{-1}[I - (K^{\circ} + (cHG)^{-1})^{-1}K^{\circ}]PA^{-1}$$

$$= + P^{-1}(HGK)^{-1}PA^{-1} .$$

Hence

$$\lim_{\substack{r \to \infty \\ c \to \infty}} s = \mu_s P^{-1}(F + \dot{\nu}) , \tag{5.12}$$

or $\mu_s^{-1}Ps = F + \dot{\nu}$, which gives

$$\dot{\hat{y}} - \dot{\nu} = F(\hat{y}, \nu) .$$

To derive the equation of the slow motions, we let μ_s equal to zero. Hence using equation (5.9) we get

$$z = - A^{-1}Dh(x) ,$$
$$\hat{y} = Pz = - PA^{-1}Dh(x) .$$

Using the fact that $- PA^{-1}D = I$, we obtain

$$\underset{\cdot}{\hat{y}} = \underset{\cdot}{y} .$$

Hence the equation of the slow motions is given by

$$\begin{aligned} \dot{x} &= f + G[K(F(\hat{y}, \nu) - \dot{\hat{y}} + \dot{\nu}) + \xi] \\ &= f + G[u + \xi] , \end{aligned}$$

where

$$u = K(F(\hat{y}, \nu) - \dot{\hat{y}} + \dot{\nu}) = K(F(y, \nu) - \dot{y} + \dot{\nu}) .$$

Remark 5.1: Note that for large but finite values of the K-matrix, the value of the control signal u remains finite (as shown in Proposition 5.1 part (b)).

INFLUENCE OF NOISE

We now investigate the influence of noise on the behavior of the system (2.1) with the control law (5.3). Assume that the output vector y is corrupted by the continuously differentiable noise r(t), thus

$$\hat{y} = y + r(t) , \tag{5.13}$$

We now find values for $\dot{y}$, $\dot{\hat{y}}$, and u. We assume that $\det(I + HGK) \neq 0$.

(a) Recall

$$\dot{y} = H[f + G[K(F(\hat{y},\nu) - \dot{\hat{y}} + \dot{\nu}(t)) + \xi]] ,$$

using equation (5.13) we get

$$\dot{y} = H[f + G[K(F - \dot{y} - \dot{r} + \dot{\nu}) + \xi]] .$$

Solving for $\dot{y}$ we obtain

$$\dot{y} = (I + HGK)^{-1}[Hf + HGKF - HGK\dot{r} + HGK\dot{\nu} + HG\xi] . \qquad (5.14)$$

(b) For the controller u,

$$\begin{aligned} u &= K(F - \dot{\hat{y}} + \dot{\nu}) \\ &= K(F - \dot{y} - \dot{r} + \dot{\nu}) , \end{aligned}$$

substituting $\dot{y} = Hf + HGu + HG\xi$, we get

$$u = K(F - Hf - HGu - HG\xi - \dot{r} + \dot{\nu}) ,$$

solving for u yields

$$u = (I + KHG)^{-1}K(F - Hf - \dot{r} - HG\xi + \dot{\nu}) . \qquad (5.15)$$

(c) The derivative of the output vector with noise is

$$\dot{\hat{y}} = Hf + HGu + HG\xi ,$$

using $u = K(F - \dot{\hat{y}} + \dot{\nu})$ we obtain

$$\dot{\hat{y}} = Hf + HGK(F - \dot{\hat{y}} + \dot{\nu}) + HG\xi ,$$

solving for $\dot{\hat{y}}$ yields

$$\dot{\hat{y}} = (I + HGK)^{-1}(Hf + HGKF + HG\xi + HGK\dot{\nu}) . \qquad (5.16)$$

In the limit the equations (5.14), (5.15), and (5.16) become

(i) $\dot{y} = F - \dot{r} + \dot{\nu}$,

(ii) $u = (HG)^{-1}[F - Hf - \dot{r} - HG\xi + \dot{\nu}]$,

(iii) $\dot{\hat{y}} - \dot{\nu} = F(\hat{y},\nu)$.

In part (i) above we can see that for an actual system, in the limiting case, the noise r(t) is fully "repeated" in the output. As for the controller u, apart from the "basic" control law $u = (HG)^{-1}(F - Hf - HG\xi + \dot{\nu})$, we have an additional component due to the additive noise.

6. CONCLUDING REMARKS

In this paper, we discussed the robustness of high-gain output feedback control designs for nonlinear time-varying uncertain models to unmodeled high-frequency dynamics. Our approach followed on the papers by Vostrikov et al. [24], and Utkin [2].

Two different control strategies were analyzed. The first one was concerned with the output regulation. To facilitate the synthesis we utilized a diffeomorphic state variable transformation of the given model into the regular form. This regular form was found very useful in the design. However the problem of constructing a transformation which brings the system into this form requires further investigation.

The aim of the second output feedback control design was to ensure the tracking by the output of a given reference signal. The proposed control algorithm involved the output vector derivative. Following Vostrikov et al. [24], we suggested a sensor estimating the output derivative. One may argue that using differentiating filters is impractical. However one has to recognize that the essential information about a given process has significant spectral components only at low frequencies [13 p. 227]. Hence if we use an approximate differentiator which is sufficiently fast then the system will hardly feel the difference between the ideal and approximate differentiators. Thus, this approximate differentiator acts as an ideal one and its gain levels off or decreases at higher frequencies. In this paper we attempted to prove that the approximate differentiator is a viable tool in the synthesis of control algorithms.

REFERENCES

[1] V. I. Utkin, *Sliding modes and their applications in variable structure systems,* Mir Publishers, Moscow, 1978.

[2] V. Utkin, "Application of equivalent control method to systems with large feedback gain," IEEE Trans. Automat. Contr., Vol. AC-23, No. 3, pp. 484-486, 1978.

[3] R. M. Hirschorn, "Invertibility of multivariable nonlinear control systems," IEEE Trans. Automat. Contr., Vol. AC-24, No. 6, pp. 855-865, 1979.

[4] R. Marino, "High-gain feedback in nonlinear control systems," Int. J. Contr., Vol. 42, No. 6, pp. 1369-1385, 1985.

[5] L. R. Hunt, R. Su, and G. Meyer, "Global transformations of nonlinear systems," IEEE Trans. Automat. Contr., Vol. AC-28, No. 1, pp. 24-31, Jan. 1983.

[6] R. Su, G. Meyer and L. R. Hunt, "Robustness in nonlinear control," in "Differential geometric control theory" Ed. R. W. Brockett et. al., pp. 316-337, Birkhäuser, Boston, 1983.

[7] A. G. Bondarev, S. A. Bondarev, N. E. Kostyleva, and V. I. Utkin, "Sliding modes in systems with asymptotic state observers," Automation and Remote Control, Vol. 46, No. 6, Part 1, pp. 679-684, June 1985.

[8] B. L. Walcott and S. H. Żak, "Output feedback control of nonlinear dynamical systems," Presented at the Int. Symp. on MTNS, Phoenix, Arizona, June 15-19, 1987.

[9] G. Leitmann, E. P. Ryan, and A. Steinberg, "Feedback control of uncertain systems: Robustness with respect to neglected actuator and sensor dynamics," Int. J. Contr., Vol. 43, No. 4, pp. 1243-1256, 1986.

[10] R. A. DeCarlo, S. H. Żak and G. P. Matthews, "Variable structure control of nonlinear multivariable systems: A tutorial," to appear, Proceedings of the IEEE, March 1988.

[11] L. R. Hunt, M. Luksic and R. Su, "Exact linearizations of input-output systems," Int. J. Control, Vol. 43, No. 1, pp. 247-255, 1986.

[12] A. Steinberg and M. Corless, "Output feedback stabilization of uncertain dynamical systems," IEEE Trans. Automat. Contr., Vol. AC-23, No. 10, pp. 1025-1027, 1985.

[13] H. M. Power and R. J. Simpson, "Introduction to dynamics and control," McGraw-Hill, England, 1978.

[14] J. Descusse and C. H. Moog, "Decoupling with dynamic compensation for strong invertible affine non-linear systems," Int. J. Control, Vol. 42, No. 6, pp. 1387-1398, 1985.

[15] A. Isidori, "The matching of a prescribed linear input-output behavior in a nonlinear system," IEEE Trans. Automat. Contr., Vol. AC-30, No. 3, pp. 258-265, 1985.

[16] L. J. Ha and E. G. Gilbert, "A complete characterization of decoupling control laws for a general class of nonlinear systems," IEEE Trans. Automat. Contr., Vol. AC-31, No. 9, pp. 823-830, 1986.

[17] M. J. Corless and G. Leitmann, "Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems," IEEE Trans. Automat. Contr., Vol. AC-26, No. 5, pp. 1139-1144, 1981.

[18] J. O'Reilly, "Robustness of linear feedback control systems to unmodelled high-frequency dynamics," Int. J. Control, Vol. 14, No. 4, pp. 1077-1088, 1986.

[19] W. T. Bauman and W. J. Rugh, "Feedback control of nonlinear systems by extended linearization," IEEE Trans. Automat. Contr., Vol. AC-31, No. 1, pp. 40-46, Jan. 1986.

[20] A. G. Lukyanov and V. I. Utkin, "Method of reducing equations for dynamic systems to a regular form," Automation and Remote Control, No. 4, pp. 5-13, April 1981.

[21] V. I. Utkin and A. S.Vostrikov, "Control systems with decoupling motions," Proc. 7th Triennial World Congress of the IFAC, Vol. 2, pp. 967-973, Helsinki, Finland, June 12-16, 1978.

[22] S. T. Glad, "Robustness of nonlinear state feedback - A survey," Automatica, Vol. 23, No. 4, pp. 425-435, 1987.

[23] S. T. Glad, "On the gain margin of nonlinear and optimal regulators," IEEE Trans. Automat. Contr., Vol. AC-29, No. 7, pp. 615-620, July 1984.

[24] A. S. Vostrikov, V. I. Utkin, and G. A. Frantsuzova, "Systems with state vector derivative in the control," Automation and Remote Contr., Vol. 43, No. 3, pp. 283-286, 1982.

[25] S. H. Żak, J. D. Brehove, and M. J. Corless, "Control of uncertain systems with unmodeled actuator and sensor dynamics and incomplete state information," to appear; IEEE Trans. Systems, Man, and Cybernetics.

[26] R. Sommer, "Control design for multivariable non-linear time-varying systems," Int. J. Control, Vol. 31, No. 5, pp. 883-891, 1980.

[27] D. Bestle and M. Zeitz, "Canonical form observer design for non-linear time-variable systems," Int. J. Control, Vol. 38, No. 2, pp. 419-431, 1983.

CONTROL OF UNCERTAIN DYNAMICAL SYSTEMS : SIMULTANEOUS STABILIZATION PROBLEMS

Bijoy K. GHOSH
Washington University
Saint-Louis, Missouri 63130, U.S.A.

In the last decade, significant progress have been witnessed in the design of a robust compensator for a family of multi input multi output systems. The main objective is to construct a dynamic compensator which simultaneously stabilizes a family of plants and satisfies various other design restrictions. The motivation is to extend various classically well-known compensator design methods for a single plant to a family of plants. Such a family of plants may occur as a result of parameter uncertainty or parameter variation in the plants and the goal is to construct a compensator which is insensitive to these parametric changes.

To begin with, we consider the "simultaneous stabilization problem" described as follows:

Given a r tuple $G_1,\dots,G_r$ of pxm proper transfer functions, does there exist a compensator $K(s)$ such that the closed loop systems $G_1[I + KG_1]^{-1},\dots, G_r[I + KG_r]^{-1}$ are internally stable?

This problem arises in reliable system design where $G_2,\dots, G_r$ represent a plant G_1 operating in various modes of failure and K(s) is a non-switching stabilizing compensator. It also arises in the stability analysis and design of a plant which can be switched into various operating modes. It has been shown in [1] that

The integer max(m,p) is the critical number of plants below which the simultaneous stabilization problem is solvable almost always i.e. generically (in a suitable topology) by a compensator of McMillan degree q_0 where q_0 is the smallest integer satisfying

$$q_0[\max(m,p) + 1-r] \geq \sum_{i=1}^{r} n_i - \max(m,p) \qquad (1)$$

In the above formula, n_i is the McMillan degree of the plant G_i for $i=1,\ldots,r$ respectively. In fact, if min(m,p) = 1 than the formula (1) also computes the minimum order of the generically stabilizing compensator. It may be remarked that the minimum order compensator problem is a classically unsolved problem and in [1] the problem is solved for the special case min(m,p) = 1.

However, beyond saying that the simultaneous stabilization problem is solvable for certain classes, it is of great interest to parameterize all those cases where the problem is indeed solvable. Moreover, for ease of computation, such a parameterization has to be explicit. This question is parameterizing the set of r tuples of plants $(G_1',\ldots,G_r')$ is addressed in [2] and one of his main results is a considerable conceptual breakthrough, since to check simultaneous stabilizability using this result one only needs to know which path component $(G_1,\ldots, G_r)$ lies in; i.e. the problem is reduced to the problem of analyzing big pieces of the space of r tuples of systems rather than individual r-tuples. Similar results on simultaneous stabilization and pole assignment for a parameterized family of plants by a parameterized family of compensators is also obtained by Dr. Ghosh and is reported in [2]. To my knowledge, use of semialgebraic geometric methods for the purpose of parameterizing stabilizable or unstabilizable path components has been done for the first time in [2].

Considering more than max(m,p) plants for the purpose of simultaneous stabilization (is quite a reasonable objective in robust system design), but unfortunately in particular in [3] it is shown that, "Pairs of simultaneously stabilizable single input single output plants of bounded McMillan degree may not have simultaneously stabilizing compensators of apriori bounded McMillan degree."

It is shown by Dr. Ghosh in [3] that there exists a sequence of pairs of simultaneously stabilizable plants of degree one for which the minimum degree of the stabilizing compensator is arbitrarily large. A consequence of the above proposition is that a simultaneously stabilizing compensator cannot be constructed by solving a set of simultaneous equations or inequalities in the coefficients of a parameterized family of compensators of a given McMillan degree. Stated differently, if r > max(m,p), the classically known algebraic and semialgebraic geometric methods are inapplicable since the compensator space is not finite dimensional and in particular, any numerical computation of the associated compensator needs to use a more appropriate transcendental method proposed by Dr. Ghosh in [4]. Also in [4] a new 'partial pole placement' problem is proposed which arises from a more practical design requirement to place an arbitrary number of self conjugate poles in the closed loop while restricting the remaining poles in the region of stability. The following result is shown:

The problem of simultaneously stabilizing three single input single output plants chosen generically is equivalent to the problem of partially pole placing one single input single output plant by a stable minimum phase compensator.

Use and application of a stable, minimum phase compensator is introduced in [4] for the first time. Furthermorea folklore example

$$\frac{s-7}{s-4.6}, \quad \frac{s-2}{2s-2.6}, \quad \frac{s-6}{4.8s-24.6}$$

of a triplet of simultaneously unstabilizable plants that are stabilizable in pairs is constructed by Dr. Ghosh [4]. These results to multi input multi output problems are further generalized in [4] to show that

"If $r \min(m,p) \leq m+p$, the simultaneous partial pole assignment problem may be analyzed via interpolation methods and one obtains a semialgebraic parameterization of the partially pole assignable r-tuples of plants. If $r \min(m,p) > m+p$, the simultaneous partial pole assignment problem is to be analyzed via transcendental methods introduced in [4]."

The above result, therefore, characterizes the "degree of difficulty" and in particular asserts the existence of certain cases (say for example $m=p$, $r\geq 3$) when interpolation methods are inapplicable in the simultaneous stabilization problem.

We have seen so far that transcendental methods are useful when the degree of the compensators under consideration is not apriori bounded. Frequently in system identification and control, it is of interest to study a family of plants for which the McMillan degree is not fixed. In particular the degree may degenerate to a lower value. Thus rather than fixing the McMillan degree of a plant, it is useful to parameterize plants of McMillan degree $\leq n$ for some n. We ,therefore, pose the following question --

"Parameterize the set Ω_n of plants of degree $\leq n$ (possibly as a semialgebraic subset of an algebraic set) such that every p in Ω_n has an open neighborhood $N(p)$ of p in Ω_n such that $N(p)$ is simultaneously stabilizable by a compensator of degree $\leq q$ for some q."

Note that this question poses robust stabilization as a parameterization problem. In [5] an explicit parameterization of Ω_n is obtained as a subset of $IRIP^{2n+1}$ for the single input single output systems and in particular we show that --

"Assume $m=p=1$, then Ω_n is a semialgebraic, open, connected and dense subset of $IRIP^{2n+1}$."

More surprisingly we show that

"Ω_n is a trivial vector bundle over a circle. In particular Ω_n is diffeomorphic to $S^1 \times IR^{2n}$."

The space Ω_n has been parameterized for a multi input multi output plant in [6] as a vector bundle over a Grassmanian, a well known object in algebraic geometry. We argue that Ω_n and not rat n (the space of strictly proper single input single output transfer functions of a given degree) or $\sum_{m,p}^{n}$ (the space of pxm transfer functions of degree n) is a more natural space for system identification and control. Various properties of this space has been reported in [8].

The geometry of Ω_n is useful in the study of a structured family of plants wherein the degree is apriori bounded. In practice, however, one is also interested in the study of a family of plants possibly with some unmodelled dynamics. For example, under the presence of a high frequency "parasitics" it is unreasonable to assume that the McMillan degree of a family of plants is bounded by n. In [6] we, therefore, construct the space Ω_∞ as a direct limit of the spaces $\Omega_1 \subset \Omega_2 \subset \ldots$ where Ω_∞ is a subspace of $\mathbb{R}^\infty$. Of course two points in Ω_∞ can model the same dynamical system and one therefore considers the quotient space $\tilde{\Omega}_\infty$ where two points in Ω_∞ are equivalent if they correspond to the same dynamical system. Various properties of $\tilde{\Omega}_\infty$ are being studied. In particular, we show that in $\tilde{\Omega}_\infty$ there exists arbitrary small open neighborhood N with the following property--

There exists a sequence $\xi_0, \xi_1, \ldots$ of plants in N such that the minimum degree of the stabilizing dynamic compensator for the plants corresponding to $\xi_0, \xi_1, \ldots$ increases arbitrarily.

This fact in particular implies that

"There exists $p \in \mathbb{R}^\infty$ such that every open neighborhood N of p in $\tilde{\Omega}_\infty$ cannot be stabilized even by an adaptive controller of arbitrary large degree q."

Thus we obtain a major limitation of the adaptive controllers that are currently of interest in system theory, viz. open neighborhoods of points in $\tilde{\Omega}_\infty$ that cannot be robustly stabilizable even by an adaptive controller. The structure of $\tilde{\Omega}_\infty$ also enables us to define a hybrid family of plants, (i.e. a family of plants with structured and unstructured uncertainty). In particular in [6] we characterize (for the first time in the literature) hybrid families of plants that can be stabilized simultaneously by an adaptive controller.

The proposed hybrid parameterization has many advantages over the currently existing graph parameterization due to Vidyasagar. In fact the hybrid parameterization is graded by the degree of the dynamical systems and each one of the graded space is diffeomorphic to an Euclidean space if the plant is strictly proper. The Euclidean structure is of particular importance in system identification. Furthermore, the sequence of plants for example

$$\delta_n(s) = \frac{s^n}{s^{n+1} + \frac{1}{n+2}}$$

converges to $\frac{1}{s}$ as $n \to \infty$ in the graph-topology. Thus in graph parameterization, arbitrary close to a plant of a given degree there exists plants of arbitrary large degree which is clearly a deficiency from the point of view of robustness and obtaining an apriori bound on the complexity of the compensators. Hybrid parameterization does not suffer from these disadvantages and therefore appears to be a good parameterization for system identification and adaptive control.

In [7] we study the problem of simultaneous stabilization of a family F of plants described as follows --

$$\begin{aligned} F \triangleq \{g(s):\ g(s) = &\ [\sum_{i=0}^{n-1} a_i s^i] \ / \ [\sum_{i=0}^{n-1} b_i s^i + s^n], \\ &a_i \in [\alpha_i, \beta_i],\ b_i \in [\gamma_i, \delta_i],\ \alpha_i \le \beta_i \\ &\gamma_i \le \delta_i,\ i = 0, \dots, n-1,\ \deg g(s) = n\} \end{aligned}$$

We prove the following rather surprising result

"A necessary and sufficient condition that every plant in F is simultaneously stabilizable by a feedback gain k is that eight plants in F (suitably chosen) is simultaneously stabilizable by a feedback gain k."

We find the above result quite surprising. Indeed it asserts the existence of a suitable family of uncountably many plants, stabilizability of which can be asserted via the simultaneous stabilization problem of a finite number of plants. This we consider is a major conceptual breakthrough.

The main idea of the preceding paragraph can be generalized to include dynamic compensation as well. In fact one can obtain a sufficient condition

which can be made asymptotically necessary by increasing the computational complexity of the algorithm. Roughly speaking one therefore concludes the existence of a computational technique to construct a robust compensator which can be asymptotically improved by considering increased computational load. This in my view is a computational breakthrough and in particular such a sequence of algorithms did not exist in the literature previously.

For the purpose of constructing a compensator with an apriori bounded McMillan degree it is important to consider to following problem.

"Given a family F of linear dynamical systems that can be stabilized simultaneously by a fixed non-switching compensator. Does there exist an apriori bound on the degree of the compensator which simultaneously stabilizes F."

In general the above problem is unsolved. However for a 1 parameter family of plant we have a surprising result: Let $x_1(s)/y_1(s)$ and $x_2(s)/y_2(s)$ be a pair of proper but not strictly proper plants. Consider a 1 parameter family F of plants described as follows

$$F = \{g_\lambda(s): \; g_\lambda(s) = [\lambda x_1 + (1-\lambda)x_2]/[\lambda y_1 + (1-\lambda)y_2]$$
$$\lambda \in [0, 1], \; \deg g_\lambda(s) \le n \; V \; \lambda\}.$$

Let $a_1, \ldots, a_t$ denote the zeros of $x_1 y_2 - x_2 y_1$ in the open left half of the complex plane. Let

$b_j = x_2/x_1(a_i)$ if the multiplicity of a_j as a common zero of x_1, x_2
is $\le$ multiplicity of a_i as a common zero of y_1, y_2

$= y_2/y_1(a_i)$ otherwise.

for $i=1, \ldots, t$. Let $s_i = (a_i-1)/(a_i+1)$ and $z_i = (\sqrt{b}_i-1)/(\sqrt{b}_i-1)$ where the branch cut for the square root is taken to be the non-positive real axis. Furthermore let k be the largest real number such that

$$[1 - k^2 z_i z_j]/[1 - s_i s_j]_{i,\, j=1}^{t}$$

is non-negative definite. The main result is now described as follows

"The following three statements are equivalent.

1. F is simultaneously stabilizable by some dynamic compensator.
2. F is simultaneously stabilizable by some dynamic compensator of degree $\leq 3n-2$.
3. $k > 1$

We find that the above result is quite surprising. In fact, where as the conjecture - "pairs of simultaneoulsy stabilizable plants of bounded McMillan degree have simultaneously stabilizing compensators of bounded McMillan degree" - is false, the conjecture that "simultaneously stabilizable linear 1-parameter family of plants of bounded McMillan degree have simultaneously stabilizing compensators of bounded McMillan degree" is indeed true. Of course it is unknown if similar results would continue to be true for multiparameter family of plants. It appears however, in view of the above result, that the problem of stabilizing a discrete r-tuple of plants (in particular a pair of plants) simultaneously is a much harder problem to solve compared to simultaneously stabilizing a continuous family of plants. This fact indeed appears to be quite contrary to our original expectation - in fact the problem of simultaneous stabilization of a pair of plants was originally used with an idea of simplifying the robust stabilization problem of a family of plants.

In order to arbitrary tune the closed loop frequencies of a plant, it is necessary to consider the simultaneous pole assignment problem. In [6] we analyze the pole placement problem as an intersection problem and apply Schubert enumerative calculus to compute (under appropriate cases) the number of complex dynamic compensators that would place the closed loop poles of a set of r-plants in a given set of self-conjugate complex numbers. We compactify the space of compensators and define a set of points known as 'base locus' and a set of points known as 'critical points.' Roughly speaking, we assert in [6] that a compensator has to avoid the base locus and the critical points for otherwise the closed loop response of the control system would either be sensitive or would fail to be robust with respect to changes in the parameters. An explicit parameterization of these points also open up some new restrictions in the compensator design problem previously unknown in system theory.

To summarize, we maintain that the use of semialgebraic geometric, algebraic geometric and transcendental methods are three distinct foundational techniques that have been applied in robust system design. Extensions of these methods to parameterization, design, identification problems, and adaptive control would be useful and are currently being explored. These techniques are also being extended to nonlinear and time varying systems.

References:

[1] B. K. Ghosh and C. I. Byrnes, "Simultaneous Stabilization and Simultaneous Pole-placement by Non-switching Dynamic Compensation," IEEE Transactions on Automatic Control, Vol. AC-28, No. 6, June 1983, pp. 735-741.

[2] B. K. Ghosh, "An Approach to Simultaneous System Design, Part I: Semialgebraic Geometric Methods," SIAM J. on Control and Optimization, May, 1986.

[3] B. K. Ghosh, "Simultaneous Partial Pole Placement--A New Approach to Multimode System Design," IEEE Trans. on Automatic Control, May, 1986.

[4] B. K. Ghosh, "Transcendental and Interpolation Methods in Simultaneous Stabilization and Simultaneous Partial Pole Placement Problem," Accepted by SIAM J. on Control and Optimization.

[5] B. K. Ghosh and W. P. Dayawansa, "A hybrid parameterization of linear single input single output systems," accepted by Systems and Control Letters.

[6] B. K. Ghosh, "An approach to simultaneous System Design, Part II: Dynamic Compensation by Algebraic Geometric Methods," (To appear in SIAM J. of Control and Optimization.)

[7] B. K. Ghosh, "Some New Results on the Simultaneous Stabilizability of a Family of Single Input, Single Output Systems," Systems and Control Letters, 6, (1985), pp. 39-45.

[8] B. K. Ghosh and W. P. Dayawansa, "An approach to linear system identification, I Differential geometric methods in hybrid parameterization problems," submitted to SIAM J. on Control and Optimization.

ROBUST MODEL TRACKING FOR A CLASS OF SINGULARLY PERTURBED NONLINEAR SYSTEMS VIA COMPOSITE CONTROL

F. Garofalo and L. Glielmo

Dipartimento di Informatica e Sistemistica
Universita' degli Studi di Napoli

1. Introduction

Typical problems encountered in the design of a control system are the presence of parameter uncertainties and the coexistence of slow and fast dynamics in the plant to be controlled. When the uncertainties are described assigning their range of variation and these variations belongs to appropriate subspaces, the so called deterministic control of uncertain systems (Leitmann, 1980; Corless-Leitmann, 1981) represents an useful tool for the design of controllers capable of guaranteeing certain performance no matter what the realization of the uncertainties is. The rigorous treatment of systems with two-time scale behavior can be done utilizing singular perturbation theory (Kokotovic *et al.*; 1986). The simultaneous use of these two methods for the control of uncertain two-time scale systems has recently received some attention (see Leitmann (this volume) and its references).

In this paper we use a composite control technique in conjunction with the robust design of controllers for uncertain systems to synthesize a nonlinear controller which forces a class of two-time scale nonlinear system to follow a two-time scale linear reference model. The controllers that are used in the two phases of the design are obtained via a constructive use of Lyapunov functions (Kalman-Bertram, 1960). The same Lyapunov functions are successively combined (as suggested by Saberi-Khalil, 1984) for obtaining the proof of ultimate boundedness of the model tracking error.

2. Problem Statement

We consider a two-time scale nonlinear system described by the following equations

$$\dot{x}(t) = A_{11}(x(t))x(t)+A_{12}(x(t))z(t)+B_1(x(t))u(t)+a_1(x(t)); \tag{2.1a}$$

$$\mu\dot{z}(t) = A_{21}(x(t))x(t)+A_{22}(x(t))z(t)+B_2(x(t))u(t)+a_2(x(t)); \tag{2.1b}$$

$$x(t_0) = x_0; \tag{2.1c}$$

$$z(t_0) = z_0; \tag{2.1d}$$

where $x(t)\in R^n$, $z(t)\in R^m$ represent the state of the system, $u(t)\in R^p$ is the control input, $a_1(x(t))$ and $a_2(x(t))$ are nonlinear vectors, $\mu\in(0,\infty)$ is the singular perturbation parameter, and $A_{ij}(\cdot)$ and $B_i(\cdot)$, $i=1,2$, $j=1,2$ are matrices of appropriate dimensions.

The reference model specifying the state behavior expected from the controlled plant is described by the linear, time-invariant system

$$\dot{\hat{x}}(t) = \hat{A}_{11}\hat{x}(t)+\hat{A}_{12}\hat{z}(t)+\hat{B}_1\hat{u}(t); \tag{2.2a}$$

$$\mu\dot{\hat{z}}(t) = \hat{A}_{21}\hat{x}(t)+\hat{A}_{22}\hat{z}(t)+\hat{B}_2\hat{u}(t); \tag{2.2b}$$

$$\hat{x}(t_0) = \hat{x}_0; \tag{2.2c}$$

$$\hat{z}(t_0) = \hat{z}_0; \tag{2.2d}$$

where $\hat{x}(t)\in R^n$ and $\hat{z}(t)\in R^m$ is the state and $\hat{u}(t)\in R^p$ is a reference signal.

The following assumptions define the class of nonlinear plants considered here.

Assumption 1. There exist full rank matrices B_i, $i=1,2$ such that, for all $x\in R^n$, the following decomposition holds:

$$B_i(x) = \bar{B}_i + \bar{B}_i E_i(x), \qquad i=1,2,$$

$$a_i(x) = \bar{B}_i d_i(x), \qquad i=1,2 \; ,$$

where $E_i(\cdot)$ (resp. $d_i(\cdot)$) is a matrix (resp. a vector) of appropriate dimensions, continuously differentiable with respect to x.

The relationship between the system (2.1) and the reference model represented by equations (2.2) is precised by the following assumptions.

Assumption 2. For all $x \in R^n$ the following equalities hold

$$A_{ij}(x) - \hat{A}_{ij} = \bar{B}_i C_{ij}(x) \; , \qquad i,j = 1,2$$

$$\hat{B}_i = \bar{B}_i \hat{C}_i \; , \qquad i = 1,2$$

where $C_{ij}(x)$ are continuously differentiable matrices.

Moreover, the singularly perturbed model is assumed in *standard form*, i.e.,

Assumption 3. Matrix $\hat{A}_{22}$ is full rank.

Defining

$$\hat{A}_0 \triangleq \hat{A}_{11} - \hat{A}_{12}\hat{A}_{22}^{-1}\hat{A}_{21} \qquad (2.3)$$

we hypothesize that

Assumption 4. The pairs $(\hat{A}_0,\bar{B}_1)$ and $(\hat{A}_{22},\bar{B}_2)$ are controllable.

Assumption 5. The matrices $A_{ij}(x)$, $B_i(x)$, $a_i(x)$, *for* $i=1,2$ and $j=1,2$, are norm bounded in R^n. In particular we define

$$h_{ij} = \sup_{x\in R^n} \|C_{ij}(x)\| ,$$

$$\kappa_i = \sup_{x\in R^n} \|E_i(x)\| ,$$

$$\nu_i = \sup_{x\in R^n} \|d_i(x)\| .$$

Moreover $\kappa_i < 1$, $i=1,2$.

Finally we make the following

Assumption 6. The input reference signals $\hat{u}(\cdot)$ are such that there exist finite positive constants

$$k_s = \sup_{t\in[t_0,\infty)} \|\hat{u}_s(t)\| ,$$

$$k_f = \sup_{t\in[t_0,\infty)} \|\hat{u}_f(t)\| ,$$

where $\hat{u}_s(t)$ and $\hat{u}_f(t)$ represent the slow and the fast time scale components of $\hat{u}(t)$ and $\hat{u}(t)\triangleq\hat{u}_s(t)+\hat{u}_f(t)$. Corresponding to these signals, there exists a positive constant $\bar{\mu}$ such that, for $\mu\in(0,\bar{\mu})$ the state variables of the reference model are uniformly bounded by known constants:

$$k_{\hat{x}} = \sup_{\substack{t\in[t_0,\infty)\\ \mu\in(0,\bar{\mu})}} \|\hat{x}(t)\| ,$$

$$k_{\hat{z}} = \sup_{\substack{t\in[t_0,\infty)\\ \mu\in(0,\bar{\mu})}} \|\hat{z}(t)\|.$$

Remark 1. Assumption 1 is the so called "matching assumption" and defines the manner in which the nonlinearities enter the plant. The equalities in Assumption 1 and 2 are the so called "model matching conditions" and determine the class of model that can be tracked by the nonlinear system under consideration.

Remark 2. System (2.1) belongs to the class of singularly perturbed nonlinear system with slow nonlinearities considered by Chow-Kokotovic (1981). Note, however, that for design purposes, it is not strictly necessary to know the nonlinearities affecting the system but only a nominal linear behavior and an evaluation of the maximum deviation from this behavior as precised in Assumption 5. The composite control design for the practical stabilization of a similar class of plants is also considered by Garofalo (to appear).

The objective of the control is to synthesize a feedback control function guaranteeing that the plant tracks the model to within a bounded neighbourhood of the zero state tracking error.[1]
The procedure we propose for the synthesis of the controller is based on the separate design of controllers guaranteeing tracking of the slow approximation and of the the boundary layer approximation of the reference model. On the basis of these control laws the composite control is constructed which guarantees tracking of the model for sufficiently small values of the singular perturbation parameter μ.

[1] *A formal definition can be found in Corless (1987) or in Appendix 1.*

3. Slow Time Scale Control

Following Kokotovic *et al.* (1986) the slow approximation of the behavior of the reference model is obtained considering $\mu=0$ in (2.2b) and substituting the resulting value for variable z in (2.2a), obtaining

$$\dot{\hat{x}}_s(t) = \hat{A}_0\hat{x}_s(t)+\hat{B}_0\hat{u}_s(t), \tag{3.1}$$

where

$$\hat{B}_0 \triangleq \hat{B}_1 - \hat{A}_{12}\hat{A}_{22}^{-1}\hat{B}_2, \tag{3.2}$$

and the subscript *s* stands for slow time-scale approximation.
In order to design the controller for tracking the slow component (3.1) of the reference model, we need an approximation of system (2.1) in the slow time scale. To this end, we assume that z variable has a nominal behavior z_n which is exactly the one that $\hat{z}$ variable takes in the reference model, that is

$$\mu\dot{z}_n(t) = \hat{A}_{21}x(t)+\hat{A}_{22}z_n(t)+\hat{B}_2\hat{u}(t). \tag{3.3}$$

Correspondingly, the approximate model of slow dynamics neglects the nominal fast transients, i.e.,[2]

$$\dot{x}_s = A_{11}(x_s)x_s+A_{12}(x_s)\bar{z}_n+B_1(x_s)u_s+a_1(x_s), \tag{3.4a}$$

$$0 = \hat{A}_{21}x_s+\hat{A}_{22}\bar{z}_n+\hat{B}_2\hat{u}_s, \tag{3.4b}$$

[2] *Sometimes, when no confusion is likely to occur, we delete the time argument of the functions.*

which gives

$$\bar{z}_n = \hat{\Gamma}(x_s,\hat{u}_s) \triangleq -\hat{A}_{22}^{-1}(\hat{A}_{21}x_s + \hat{B}_2\hat{u}_s) \tag{3.5a}$$

and

$$\dot{x}_s = A_0(x_s)x_s+B_1(x_s)u_s-A_{12}(x_s)\hat{A}_{22}^{-1}\hat{B}_2\hat{u}_s+a_1(x_s), \tag{3.5b}$$

with

$$A_0(x_s) \triangleq A_{11}(x_s) - A_{12}(x_s)\hat{A}_{22}^{-1}\hat{A}_{21}. \tag{3.6}$$

Define now the slow time scale tracking error as

$$\xi_s \triangleq x_s-\hat{x}_s \ . \tag{3.7}$$

On the basis of (3.1), (3.2), (3.5) and (3.6) the slow time scale tracking error dynamics can be written as

$$\begin{aligned}\dot{\xi}_s = F_s\xi_s + \bar{B}_1u_s + \bar{B}_1E_1(x_s)u_s + \bar{B}_1[H_1(x_s)+K_s]\xi_s + \\ + \bar{B}_1[H_1(x_s)\hat{x}_s-H_2(x_s)\hat{u}_s+d_1(x_s)],\end{aligned} \tag{3.8}$$

where $F_s \triangleq \hat{A}_0-\bar{B}_1K_s$, $K_s \in R^{pxm}$ is a matrix which makes matrix F_s asymptotically stable with specified eigenvalues (which is always possible by virtue of Assumption 4), and

$$H_1(x_s) \triangleq [C_{11}(x_s)-C_{12}(x_s)\hat{A}_{22}^{-1}\hat{A}_{21}], \tag{3.9a}$$

$$H_2(x_s) \triangleq [\hat{C}_1+C_{12}(x_s)\hat{A}_{22}^{-1}\hat{B}_2]. \tag{3.9b}$$

From the knowledge of matrices $C_{11}(x)$ and $C_{12}(x)$ (given in Assumption 2), and matrices $\hat{A}_{22}$, $\hat{A}_{21}$ and $\hat{B}_2$, we can compute the following constants

$$k_{\xi_s} \triangleq \sup_{x\in R^n} \|H_1(x)+K_s\|, \qquad (3.10a)^3$$

$$k_{d_1} \triangleq \sup_{\substack{x\in R^n \\ t\in[t_0,\infty) \\ \mu\in(0,\bar{\mu})}} \|H\ (x)\hat{x}_s - H_2(x)\hat{u}_s + d_1(x)\|. \qquad (3.10b)$$

Consider now the nonlinear feedback control law (Ambrosino-Celentano-Garofalo, 1985; Garofalo-Glielmo, to appear)

$$p_s(\xi_s) \triangleq -\gamma_s \bar{B}_1^T P_s \xi_s, \qquad (3.11a)$$

where P_s is the solution of the Lyapunov equation

$$F_s^T P_s + P_s F_s = -Q_s, \qquad Q_s \text{ positive definite}, \qquad (3.11b)$$

and

$$\gamma_s = \gamma_s(\|\xi_s\|) \triangleq \frac{\gamma_{s1}+\gamma_{s2}\|\xi_s\|}{\|B_1^T P_s \xi_s\| + \delta_s}, \quad \delta_s>0. \qquad (3.11c)$$

This feedback control has the tracking capabilities described in the next theorem.

Theorem 1. Consider the slow approximation (3.4) of system (2.1) subject to the feedback control law in (3.11). If constants γ_{si}, i=1,2, in (3.11c) are chosen so as to satisfy

$$\gamma_{s1} \geq \frac{k_{d_1}}{1-\kappa_1}, \qquad (3.12a)$$

[3] *Notice that the suprema can always be replaced by upper bounds.*

$$\gamma_{s2} \geq \frac{k_{\xi_s}}{1-\kappa_1}, \tag{3.12b}$$

then system (3.5b) tracks the slow approximation (3.1) of the reference model (2.2) to within a spherical neighbourhood of $\xi_s=0$ whose radius can be made arbitrarily small by a suitable selections of constants γ_{si}, i=1,2, and/or of constant δ_s in (3.11c).

Proof. The proof of the theorem can be found in Appendix 2.

4. Fast Time Scale Control

The boundary layer approximation of the reference model (2.2) is given by (Kokotovic *et al.*, *1986)*

$$\frac{d\hat{z}_f}{d\tau} = \hat{A}_{22}\hat{z}_f + \hat{B}_2\hat{u}_f, \tag{4.1}$$

where $\tau=t/\mu$, $\hat{u}_f$ represents the fast component of the reference signal, and

$$\hat{z}_f \triangleq \hat{z} - \hat{\Gamma}(\hat{x},\hat{u}_s) = \hat{z} + \hat{A}_{22}^{-1}(\hat{A}_{21}\hat{x}+\hat{B}_2\hat{u}_s). \tag{4.2}$$

The fast time scale approximation of system (2.1) is obtained substituting the slow control expression (3.11a) in equation (2.1b) and approximating variable $x_s(t)$ by $x(t)$ and $\hat{x}_s(t)$ by $\hat{x}(t)$. So doing we obtain

$$\mu\dot{z} = A_{21}(x)x-\gamma_s B_2(x)\overline{B}_1^T P_s\xi + A_{22}(x)z + B_2(x)u_f+a_2(x), \tag{4.3}$$

where u_f is the fast component of the control law and $\xi \triangleq x-\hat{x}$.

Defining

$$z_f \triangleq z - \hat{\Gamma}(x,\hat{u}_s) = z + \hat{A}_{22}^{-1}(\hat{A}_{21}x+\hat{B}_2\hat{u}_s), \tag{4.4}$$

the boundary layer model of the system can be written as

$$\begin{aligned}\frac{dz_f}{d\tau} = \; & A_{22}(x)z_f+B_2(x)u_f-\gamma_s B_2(x)\bar{B}_1^T P_s \xi \\ & +\bar{B}_2 G_1(x)x+\bar{B}_2 G_2(x)\hat{u}_s+\bar{B}_2 d_2(x),\end{aligned} \tag{4.5}$$

with

$$G_1(x) \triangleq [C_{21}(x) - C_{22}(x)\hat{A}_{22}^{-1}\hat{A}_{21}], \tag{4.6a}$$

$$G_2(x) \triangleq -[\hat{C}_2+C_{22}(x)\hat{A}_{22}^{-1}\hat{B}_2]. \tag{4.6b}$$

The fast time scale tracking error can be defined as

$$\zeta_f \triangleq z_f-\hat{z}_f \;, \tag{4.7}$$

and, on the basis of (4.1) and (4.5), its dynamics can be written as

$$\begin{aligned}\frac{d\zeta_f}{d\tau} = \; & F_f\zeta_f + \bar{B}_2u_f + \bar{B}_2E_2(x)u_f + \bar{B}_2[C_{22}(x) + K_f]\zeta_f + \\ & + \bar{B}_2[G_1(x) - \gamma_s(I_p + E_2(x))\bar{B}_1^T P_s]\xi \\ & + \bar{B}_2[G_1(x)\hat{x} + C_{22}(x)\hat{z}_f + G_2(x)\hat{u}_s - \hat{C}_2\hat{u}_f + d_2(x)],\end{aligned} \tag{4.8}$$

where $F_f \triangleq \hat{A}_{22}-\bar{B}_2K_f$ and $K_f \in R^{pxm}$ is a matrix which makes matrix F_f asymptotically stable with specified eigenvalues (see Assumption 4).

On the basis of Assumptions 1, 2, 5 and 6, we can evaluate the finite constants

$$k_{d_2} \triangleq \sup_{\substack{x \in R^n \\ t \in [t_0, \infty) \\ \mu \in (0, \tilde{\mu})}} \| G_1(x)\hat{x} + C_{22}(x)\hat{z}_f + G_2(x)\hat{u}_s - \hat{C}_2\hat{u}_f + d_2(x) \|, \tag{4.9a}$$

$$k_\zeta \triangleq \sup_{x \in R^n} \| C_{22}(x) + K_f \|, \tag{4.9b}$$

$$k_\xi \triangleq \sup_{x \in R^n} \| G_1(x) - \gamma_s (I_p + E_2(x)) \bar{B}_1^T P_s \|, \tag{4.9c}$$

In the fast time scale the variables x and ξ can be considered constants, and the fast control law we propose for making the boundary layer system track the boundary layer reference model has the form

$$p_f(\zeta_f) \triangleq -\gamma_f \bar{B}_2^T P_f \zeta_f, \tag{4.10a}$$

where P_f is the solution of the Lyapunov equation

$$F_f^T P_f + P_f F_f = -Q_f, \qquad Q_f \text{ positive definite}, \tag{4.10b}$$

and

$$\gamma_f = \gamma_f(\|\zeta_f\|, \|\xi\|) \triangleq \frac{\gamma_{f1} + \gamma_{f2}\|\zeta_f\| + \gamma_{f3}\|\xi\|}{\|\bar{B}_2^T P_f \zeta_f\| + \delta_f}, \quad \delta_f > 0. \tag{4.10c}$$

We can state the following

Theorem 2. Consider the boundary layer approximation (4.5) of system (2.1) subject to the feedback control law (4.10). If constants γ_{fi}, i=1,...,3 are chosen so as to satisfy

$$\gamma_{f1} \geq \frac{k_{d_2}}{1-\kappa_1}\ , \tag{4.11a}$$

$$\gamma_{f2} \geq \frac{k_{\zeta}}{1-\kappa_1}\ , \tag{4.11b}$$

$$\gamma_{f3} \geq \frac{k_{\xi}}{1-\kappa_1}\ , \tag{4.11c}$$

then system (4.5) tracks the boundary layer *reference* model (4.1) to within a spherical neigbourhood of $\zeta_f=0$ whose radius can be made arbitrarily small by a suitable selection of constants γ_{fi}, $i=1,\ldots,3$ and/or constant δ_f in (4.10c).

Proof. The proof can be found in Appendix 2.

5. Guaranteed Performance of the Composite Control

The composite control is obtained as the sum of the slow and the fast control law with variable ζ_f replaced by $\zeta-[\hat{\Gamma}(x,\hat{u})+\hat{\Gamma}(\hat{x},\hat{u})] = \zeta + \hat{A}_{22}^{-1}\hat{A}_{21}\xi$, and ξ_s by its approximation ξ, obtaining

$$u_c = -\gamma_s\overline{B}_1^T P_s\xi - \gamma_f\overline{B}_2^T P_f\zeta - \gamma_f\overline{B}_2^T P_f\hat{A}_{22}^{-1}\hat{A}_{21}\xi, \tag{5.1}$$

where $\zeta \triangleq (z-\hat{z})$.

For this control law we can establish the following theorem.

Theorem 3. Consider system (2.1) subject to the control law (5.1). The closed loop system tracks the reference model to within a spherical neighbourhood of the zero state tracking error, if the following conditions are satisfied.

i) The constant γ_{s1} satisfies the inequality:

$$\gamma_{s1} \geq \frac{k_{d_1}'}{1 - \kappa_1} \ , \tag{5.2}$$

with

$$k_{d_1}' \triangleq \sup_{\substack{x\in R^n \\ t\in[t_0,\infty) \\ \mu\in(0,\bar{\mu})}} \|H(x)\hat{x} - H_2(x)\hat{u} + d_1(x)\| , \tag{5.3}$$

and the constant γ_{s2} satisfies the inequality (3.12b);

ii) the constants γ_{fj}, j=1,2,3 in the control law (5.1) are chosen so as to satisfy inequalities (4.11);

iii) constant γ_{f3}, besides satisfying (4.11c), satisfies

$$\gamma_{f3} < \frac{\lambda_{min}(Q_s)}{2\|P_s\| \sup_{x\in R^n} \|B_1(x)\|} \tag{5.4}$$

iv) the singular perturbation parameter is such that $0<\mu<\mu^*$ where μ^* is a constant whose value can be *a priori* computed.

Proof. The proof of Theorem 3 and the expression for the upper bound of paramerer μ are given in Appendix 3.

6. Conclusions

The robust model tracking control presented here is designed using the approach of deterministic control of uncertain systems, together with the composite control technique developed for singularly perturbed systems. This enables the designer to guarantee the model following within a spherical neighbourhood of the zero error, in the presence of "slow" nonlinearities. It must be pointed out that this technique does not require the knowledge of the form of the nonlinearities, but just the possible range of their variations.

Appendix 1

Some definitions and a useful lemma.

Consider the equation of a model tracking error dynamics in the form

$$\dot{\epsilon} = \varphi(\epsilon,t) \ , \ \epsilon(t_0)=\epsilon_0, \tag{A1.1}$$

where $t \in R$, $\epsilon \in R^p$, and $\varphi : R^p x R \rightarrow R^p$ We say that the system tracks the reference model to within a spherical neighbourhood of radius R of $\epsilon=0$ (indicated with B(R)) iff the following properties are satisfied:

i) *Existence of the solution*. Given any $(\epsilon_0, t_0) \in R^p x R$ there exists a solution $\epsilon(\cdot):[t_0,t_1) \rightarrow R^p$, $t_1 > t_0$ of (A1.1).

ii) *Indefinite extension of solution*. Every solution $\epsilon(\cdot):[t_0,t_1) \rightarrow R^p$ of (A1.1) has an extension over $[t_0,\infty)$.

iii) *Global uniform boundedness*. Given any bound $r \in R_+$, there exists a bound $d(r) \in R_+$ such that if $\epsilon(\cdot):[t_0,t_1) \rightarrow R^p$ is a solution of (A1.1) with $\|\epsilon_0\| \leq r$, then $\|\epsilon(t)\| \leq d(r)$ for all $t \in [t_0,t_1)$.

iv) *Local boundedness within* B(R). There exists a spherical neighbourhood $B(R_0)$ of $\epsilon=0$ such that if $\epsilon(\cdot):[t_0,t_1) \rightarrow R^p$ is a solution of (A1.1) with $\epsilon_0 \in B(R_0)$ then $\epsilon(t) \in B(R)$ for all $t \in [t_0,t_1)$.

v) *Global uniform ultimate boundedness within* B(R). Given any bound $r \in R_+$ there exists $T(r) \in R_+$ such that if $\epsilon(\cdot):[t_0,t_1) \rightarrow R^p$ is a solution of (A1.1) with $\|\epsilon_0\| \leq r$, then $\epsilon(t) \in B(R)$ for all $t \geq t_0 + T(r)$.

The listed properties of the solution $\epsilon(\cdot):[t_0,t_1) \rightarrow R^p$ can be stated with the aid of the following lemma (for the proof see Corless-Leitmann, 1981).

Lemma. Given system (A1.1) suppose $\varphi(0,t)=0$ for all $t\in\mathbb{R}$. If there exists a C^1 function L defined on $\|\epsilon\|\geq s$ and $t\in\mathbb{R}$, and if there exist class $\mathbb{KR}$ functions χ_1 and χ_2 and a class $\mathbb{K}$ function χ_3 such that

$$\chi_1(\|\epsilon\|)\leq L(\epsilon,t)\leq\chi_2(\|\epsilon\|), \tag{A1.2a}$$

$$\frac{\partial}{\partial t}L(\epsilon,t)+\nabla_\epsilon^T L(\epsilon,t) \leq -\chi_3(\|\epsilon\|), \tag{A1.2b}$$

then for all $\|\epsilon\|\geq s$ and $t\in\mathbb{R}$ the system tracks the reference model to within any spherical neighbourhood $B(\bar{R})$ of $\epsilon=0$ with $\bar{R}>\chi_1^{-1}\circ\chi_2(s)$.

Appendix 2

Proofs of Theorems 1 and 2.

Consider as Lyapunov function candidate for system (3.8) with the feedback control (3.11) the following

$$V(\xi_s) \triangleq \xi_s^T P_s \xi_s . \tag{A2.1}$$

Evaluating the derivative along the solutions of the closed loop system by virtue of (3.9), (3.10), (3.11), (3.12) and Assumption 5, we have

$$\begin{aligned}(1/2)\dot{V}(\xi_s) = & -(1/2)\xi_s^T Q_s \xi_s - \gamma_s \xi_s^T P_s \bar{B}_1 \bar{B}_1^T P_s \xi_s \\ & - \gamma_s \xi_s^T P_s \bar{B}_1 E_1(x_s) \bar{B}_1^T P_s \dot{\xi}_s + \xi_s^T P_s \bar{B}_1 H_1(x_s) x_s \\ & - \xi_s^T P_s \bar{B}_1 H_2(x_s) \hat{u}_s + \xi_s^T P_s \bar{B}_1 d_1(x_s) + \xi_s^T P_s \bar{B}_1 K_s \xi\end{aligned}$$

$$\begin{aligned}\leq & -(1/2)\xi_s^T Q_s \xi_s - \gamma_s \|\bar{B}_1^T P_s \xi_s\|^2(1-\kappa_1) + \|\bar{B}_1^T P_s \xi_s\| \|H_1(x_s)+K_s\| \|\xi_s\| \\ & + \|\bar{B}_1^T P_s \xi_s\| \|H_1(x_s)\hat{x}_s - H_2(x_s)\hat{u}_s + d_1(x_s)\|\end{aligned}$$

$$\begin{aligned}\le & -(1/2)\xi_s^T Q_s \xi_s - (\gamma_{s1}+\gamma_{s2}\|\xi_s\|)(\|\overline{B}_1^T P_s \xi_s\| - \delta_S)(1-\kappa_S) \\ & + \|\overline{B}_1^T P_s \xi_s\| \|H_1(x_s)+K_s\| \|\xi_s\| \\ & + \|\overline{B}_1^T P_s \xi_s\| \|H_1(x_s)\hat{x}_s - H_2(x_s)\hat{u}_s + d_1(x_s)\|\end{aligned}$$

$$\le -(1/2)\xi_s^T Q_s \xi_s + k_{d_1}\delta_s + k_{\xi_s}\delta_s\|\xi_s\|$$

$$\le (1/2)[-v_1\|\xi_s\|^2 + v_2\|\xi_s\| + v_3] \qquad (A2.2)$$

where $v_1 \triangleq \lambda_{min}(Q_s)$, $v_2 \triangleq 2k_{\xi_s}\delta_s$, and $v_3 \triangleq 2k_{d_1}\delta_s$.

At this stage the application of the lemma reported in Appendix 1 proves the statement of the Theorem 1.

The proof of Theorem 2 proceeds exactly in the same way. We define as Lyapunov candidate for system (4.5) subject to the feedback control (4.10)

$$W(\zeta_f) \triangleq \zeta_f^T P_f \zeta_f . \qquad (A2.3)$$

The derivative along the solutions of the closed loop system, considering x constant in the fast time scale, can be proved to satisfy the following inequality

$$\dot{W}(\zeta_f) \le -w_1\|\zeta_f\|^2 + w_2\|\zeta_f\| + w_3 \qquad (A2.4)$$

with $w_1 \triangleq \lambda_{min}(Q_f)$, $w_2 \triangleq 2\delta_f k_\xi$, and $w_3 \triangleq w_3' + w_3''\|\xi\| \triangleq 2\delta_f k_{d_2} + 2\delta_f k_\xi\|\xi\|$.

Appendix 3

Proof of Theorem 3.

The proof of the theorem is based on the combined use of two Lyapunov functions, one for each component of the model reference tracking error.

For the first component we can write

$$
\begin{aligned}
\dot{\xi} = {} & A_{11}(x)x+A_{12}(x)z-\gamma_s B_1(x)\bar{B}_1^T P_s\xi-\gamma_f B_1(x)\bar{B}_2^T P_f\zeta \\
& + \gamma_f B_1(x)\bar{B}_2^T P_f\hat{A}_{22}^{-1}\hat{A}_{21}\xi+a_1(x)-\hat{A}_{11}\hat{x}-\hat{A}_{12}\hat{z}-\hat{B}_1\hat{u} \\
= {} & \{F_s\xi-\gamma_s\bar{B}_1\bar{B}_1^T P_s\xi-\gamma_s\bar{B}_1 E_1(x)\bar{B}_1^T P_s\xi+\bar{B}_1[H_1(x)+K_s]\xi+ \\
& +\bar{B}_1[H_1(x)\hat{x}-H_2(x)\hat{u}+d_1(x)]\}\ -\gamma_f B_1(x)\bar{B}_2^T P_f[\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi] \\
& +A_{12}(x)[\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi]+\bar{B}_1 C_{12}(x)[\hat{z}-\hat{\Gamma}(\hat{x},\hat{u})] \qquad \text{(A3.1)}
\end{aligned}
$$

The terms within braces are exactly the same as in the slow model (3.8), taking apart the substitution of $\hat{x}_s$ and $\hat{u}_s$ with $\hat{x}$ and $\hat{u}$. On the basis of Assumptions 5 and 6, and recalling (4.10c), it is possible to find constants α_i, $i=1,3$ and a_i, $i=1,3$ such that

$$
\begin{aligned}
& \|-\gamma_f B_1(x)\bar{B}_2^T P_f[\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi] \\
& \quad +A_{12}(x)[\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi]-\bar{B}_1 C_{12}(x)[\hat{z}-\hat{\Gamma}(\hat{x},\hat{u})]\| \\
& \le \alpha_1\|\xi\|+\alpha_2\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\|+\alpha_3, \qquad \text{(A3.2a)}
\end{aligned}
$$

$$
\begin{aligned}
\|\dot{\xi}\| = {} & \|A_{11}(x)x+A_{12}(x)z-\gamma_s B_1(x)\bar{B}_1^T P_s\xi-\gamma_f B_1(x)\bar{B}_2^T P_f\zeta \\
& - \gamma_f B_1(x)\bar{B}_2^T P_f\hat{A}_{22}^{-1}\hat{A}_{21}\xi+a_1(x)-\hat{A}_{11}\hat{x}-\hat{A}_{12}\hat{z}-\hat{B}_1\hat{u}\| \\
\le {} & a_1'\|\xi\|+a_2'\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\|+a_3'. \qquad \text{(A3.2b)}
\end{aligned}
$$

For the second component of the model tracking error we simply rewrite equation (4.8) as

$$
\begin{aligned}
\mu\dot{\zeta} \;=\;& [F_f-\gamma_{f2}B_2(x)\overline{B}_2^T P_f](\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi) \\
&+ \overline{B}_2[C_{22}(x)+K_f](\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi) \\
&+ \overline{B}_2[G_1(x)-\gamma_s(I_p+E_2(x))\overline{B}_1^T P_s]\xi \\
&+ \overline{B}_2[G_1(x)\hat{x}+C_{22}(x)\hat{z}_f+G_2(x)\hat{u}_s-\hat{C}_2\hat{u}_f+d_2(x)]. \qquad \text{(A3.3)}
\end{aligned}
$$

Consider the function

$$
W(\xi,\zeta) \triangleq 1/2\ (\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi)^T P_f(\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi), \qquad \text{(A3.4)}
$$

and evaluate the derivative along the solutions of the closed loop tracking error system (A3.1), (A3.3). One obtains

$$
\begin{aligned}
\dot{W}(\xi,\zeta) =& -(\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi)^T P_f\hat{A}_{22}^{-1}\hat{A}_{21}\dot{\xi} + (\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi)^T P_f\dot{\zeta} \\
&\le \|P_f\hat{A}_{22}^{-1}\hat{A}_{21}\|\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\|[a_1'\|\xi\|+a_2'\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\|+a_3'\ ] \\
&\quad + \frac{1}{\mu}\ [-w_1\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\|^2 + w_2\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\|] \\
&\triangleq -(\frac{w_1}{\mu} - a_2)\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\|^2 + a_1\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\|\|\xi\| \\
&\quad +(a_3+\frac{w_2}{\mu})\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\| + \frac{w_3'}{\mu} + \frac{w_3''}{\mu}\|\xi\|. \qquad \text{(A3.5)}
\end{aligned}
$$

Consider now the function

$$
V(\xi) \triangleq 1/2\ \xi^T P_s\xi, \qquad \text{(A3.6)}
$$

and evaluate the time derivative along the solutions of the closed loop tracking error system (A3.1), (A3.3). In view of (A3.2a), and conditions (5.2) and (5.3), we have

$$\dot{V}(\xi) = -v_1\|\xi\|^2+v_2\|\xi\|+v_3+2\|P_s\|\|\xi\|[\alpha_1\|\xi\|+\alpha_2\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\|+\alpha_3]$$

$$\triangleq -(v_1-2\alpha_1\|P_s\|)\|\xi\|^2 + b_1\|\xi\| + b_2\|\xi\|\|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\| + v_3. \tag{A3.7}$$

We can choose as Lyapunov candidate for the closed loop tracking error system (A3.1), (A3.3) the following

$$L(\xi,\zeta) \triangleq \begin{bmatrix} \xi \\ \zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi \end{bmatrix}^T P(c) \begin{bmatrix} \xi \\ \zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi \end{bmatrix}, \tag{A3.8}$$

where

$$P(c) \triangleq \begin{bmatrix} (1-c)P_s & 0 \\ 0 & cP_f \end{bmatrix}, \quad 0\le c\le 1. \tag{A3.9}$$

In view of (A3.5) and (A3.7) the time derivative of (A3.8) along the solutions of the closed loop tracking error system satisfies

$$\dot{L}(\xi,\zeta) \le - \begin{bmatrix} \|\xi\| \\ \|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\| \end{bmatrix}^T M(c) \begin{bmatrix} \|\xi\| \\ \|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\| \end{bmatrix}$$

$$+ m^T(c) \begin{bmatrix} \|\xi\| \\ \|\zeta+\hat{A}_{22}^{-1}\hat{A}_{21}\xi\| \end{bmatrix} + \bar{m}, \tag{A3.10}$$

where

$$M(c) \triangleq \begin{bmatrix} (1-c)(v_1-2\alpha_1\|P_s\|) & -1/2(ca_1+(1-c)b_2 \\ -1/2(ca_1+ (1-c)b_2 & c(\frac{w}{\mu}1- a_2) \end{bmatrix}, \tag{A3.11a}$$

$$m(c) \triangleq [(1-c)b_1+c\frac{w''_3}{\mu} \qquad c(a_3+\frac{w_2}{\mu})], \tag{A3.11b}$$

and

$$\bar{m} \triangleq v_3(1-c) + c\frac{w'_3}{\mu} . \tag{A3.11c}$$

Provided that $\alpha_1<\frac{v_1}{2\|P_s\|}$ (which is guaranteed by condition (5.4)), the upper bound μ_p *of parameter* μ which guarantees the definite positivity of matrix M(c) is given by (see Saberi-Khalil, 1984)

$$\mu_p = \frac{(v_1-2\alpha_1\|P_s\|)w_1}{(v_1-2\alpha_1\|P_s\|)a_2 + a_1b_2} \tag{A3.12}$$

Chosen $\mu^* \triangleq \min(\bar{\mu},\mu_p)$, for each $0<\mu<\mu^*$ the application of the lemma contained in Appendix 1 completes the proof of the Theorem.

References

G. AMBROSINO, G. CELENTANO, F. GAROFALO

(1985) Robust model tracking control for a class of nonlinear plants, IEEE Trans. Automat. Contr. AC-30, pp. 275-279.

J.H. CHOW, P.V. KOKOTOVIC

(1981) A two-stage Lyapunov-Bellman feedback design of a class of nonlinear systems, *IEEE Trans* on Automat. Control AC-26, pp. 656-663.

M. CORLESS

(1987) Robustness of a class of feedback-controlled uncertain nonlinear systems in the presence of singular perturbations, Proc. of the ACC Conf., Minneapolis, MI.

M. CORLESS, G. LEITMANN

(1981) Continuous state feedback guaranteeing uniform ultimate boundedness for uncertain dynamic systems, IEEE Trans. Automat. Contr. AC-26, pp.1139-1144.

F. GAROFALO

(to appear) Composite control of a singularly perturbed uncertain system with slow nonlinearities, Int. J. of Control.

F. GAROFALO, L. GLIELMO

(to appear) Nonlinear continuous feedback control for robust tracking, in "Deterministic Control of Uncertain Systems", IEE Press: London.

R.E. KALMAN, J.E. BERTRAM

(1960) Control system analysis and design via the second method of Lyapunov: Continuous-time systems , ASME Journal of Basic Engineering, pp. 371-393.

P.V. KOKOTOVIC, H.K. KHALIL, J. O'REILLY

(1986) "Singular Perturbation Methods in Control: Analysis and Design", Academic Press: London.

G. LEITMANN

(1980) Deterministic control of uncertain systems, Astronautica Acta 7, pp. 1457-1461.

(this volume) Controlling singularly perturbed uncertain dynamical systems.

A. SABERI, H. KHALIL

(1984) Quadratic-type Lyapunov functions for singularly perturbed systems, IEEE Trans. Automat. Contr. AC-29, pp. 542-550.

MODELISATION ET COMMANDE EN ECONOMIE

MODELS AND CONTROL POLICIES IN ECONOMICS

Qualitative Differential Games: a Viability Approach

Jean-Pierre Aubin
CEREMADE, Université de Paris-Dauphine
75775 PARIS cx(16)

March 13, 1988

Abstract

We define the playability property of a qualitative differential game, and we *characterize it* by a regulation map which associates with any playable state a set of playable controls. We extract among theses playable controls the set of discriminating and pure controls of one of the players. We characterize them through an adequate "contingent" Hamilton-Jacobi-Isaacs equation, and we provide sufficient conditions implying the existence of continuous or minimal playable, discriminating and pure feedbacks.

Résumé

Nous définissons une propriété de jouabilité de jeux différentiels qualitatifs, que nous caractérisons à l'aide d'une correspondance de régulation qui associe à tout état jouable un ensemble de contrôles jouables. On distingue parmi ces contrôles jouables l'ensemble des contrôles discriminants et des contrôles purs d'un des joueurs. Nous caratérisons ces concepts par une équation d'Hamilton-Jacobi-Isaacs "contingente", et nous énonçons des conditions suffisantes impliquant l'existence de retroactions jouables, discriminantes et pures.

We consider a two-player differential game whose dynamics are described by

$$\left\{\begin{array}{ll} a) & \left\{\begin{array}{llll} i) & x'(t) & = & f(x(t),y(t),u(t)) \\ ii) & u(t) & \in & U(x(t),y(t)) \end{array}\right. \\ b) & \left\{\begin{array}{llll} i) & y'(t) & = & g(x(t),y(t),v(t)) \\ ii) & v(t) & \in & V(x(t),y(t)) \end{array}\right. \end{array}\right.$$

The rules of the game are set-valued maps $P : Y \rightsquigarrow X$ and $Q : X \rightsquigarrow Y$, stating the constraints imposed by one player on the other.

The **playability domain** of the game $K \subset X \times Y$ is defined by:

$$K \ := \ \{ \, (x,y) \in X \times Y \ | \ x \in P(y) \ \text{ and } \ y \in Q(x) \, \}$$

(We consider only the time-independent case for the sake of simplicity).

The **playability property** states that for all initial state $(x_0, y_0) \in K$, there exists a solution to the differential game which is playable in the sense that

$$\forall\, t \geq 0, \quad x(t) \ \in \ P(y(t)) \ \ \& \ \ y(t) \ \in \ Q(x(t))$$

We shall charaterize it by constructing **decision rules**

$$(x,y,v) \rightsquigarrow \Phi(x,y;v) \ \ \& \ \ (x,y,u) \rightsquigarrow \Psi(x,y;u)$$

which involve the contingent derivatives[1] of the set-valued maps P and Q, with which we build the **regulation map** R mapping each $(x,y) \in K$ to the regulation set

$$R(x,y) \ = \ \{ \, (u,v) \,|\, u \in \Phi(x,y;v) \ \text{ and } \ v \in \Psi(x,y;u) \, \}$$

The controls belonging to $R(x,y)$ are called **playable**.

[1] We recall that the contingent cone $T_K(x)$ to a subset K at $x \in K$ is the closed cone of elements v satisfying

$$\liminf_{h \to 0+} d(x + hv, K)/h \ = \ 0$$

The contingent derivative of the set-valued map Q from X to Y at a point (x,y) of its graph is the closed positively homogenous set-valued map $DQ(x,y)$ from X to Y defined by

$$\text{Graph}(DQ(x,y)) \ := \ T_{\text{Graph}(Q)}(x,y)$$

or, equivalently, by.

$$v \in DQ(x,y)(v) \iff \liminf_{h \to 0+,\, u' \to u} d\left(v, \frac{Q(x+hu') - y}{h}\right) = 0$$

The Playability Theorem states that under technical assumptions, the playability property holds true if and only if

$$\forall\,(x,y)\in K,\quad R(x,y)\ \neq\ \emptyset$$

and that playable solutions to the game are regulated by the regulation law:

$$\forall\, t\geq 0,\quad u(t)\ \in\ \Phi(x(t),y(t);v(t))\ \ \&\ \ v(t)\ \in\ \Psi(x(t),y(t);u(t))$$

We then introduce the subset

$$A(x,y;v)\ :=\ \{\ u\in U(x,y)\mid (u,v)\in R(x,y)\ \}$$

of discriminating controls which allow the first player to associate to any control $v\in V(x,y)$ played by the second player at least a control $u\in U(x,y)$ such that the pair (u,v) is playable and the subset

$$B(x,y)\ :=\ \bigcap_{v\in V(x,y)} A(x,y;v)$$

of pure controls which allow the first player to find a control $u\in U(x,y)$ such that (u,v) is playable for all $v\in V(x,y)$.

These concepts are particularly relevant for games "against nature" or "disturbances"(see [11,12,26,27] and their references).

Before going further, it may be useful to relate these concepts to more familiar ones through an adequate Hamilton-Jacobi-Isaacs's equation (see[18]).

For that purpose, we characterize the rules P and Q by their indicator functions W_P and W_Q defined respectively by

$$W_P(x,y)\ :=\ \begin{cases} 0 & \text{if } x\in P(y)\\ +\infty & \text{if } x\notin P(y)\end{cases}\qquad W_Q(x,y)\ :=\ \begin{cases} 0 & \text{if } y\in Q(x)\\ +\infty & \text{if } y\notin Q(x)\end{cases}$$

These functions are only lower semicontinuous, but we can still "differentiate"them by taking their contingent epiderivatives[2]. We set

$$H(W_P+W_Q;x,y;u,v)\ :=\ D_\uparrow(W_P+W_Q)(x,y)(f(x,y;u),g(x,y;v))$$

[2] The contingent derivative $D_\uparrow W(x)$ of a extended function W from X to $\mathbf{R}\cup\{+\infty\}$ at $x\in\mathrm{Dom}(W)$ is defined by

$$\mathcal{E}pD_\uparrow W(x)\ :=\ T_{\mathcal{E}p(W)}(x,W(x))$$

or, equivalently, by

$$D_\uparrow W(x)(u)\ =\ \liminf_{h\to 0+,u'\to u}\frac{V(x+hu')-V(x)}{h}$$

We shall prove that

— the game is playable if and only if

$$\inf_{u\in U(x,y),v\in V(x,y)} H(W_P + W_Q; x, y; u, v) \; = \; 0$$

and the regulation map is equal to

$$\left\{ \begin{array}{l} R(x,y) = \{(u,v) \in U(x,y) \times V(x,y) \mid \\ H(W_P + W_Q; x, y; u, v) = \inf_{u'\in U(x,y),v'\in V(x,y)} H(W_P + W_Q; x, y; u', v')\} \end{array} \right.$$

— the first player has a discriminating control if and only if

$$\sup_{v\in V(x,y)} \inf_{u\in U(x,y)} H(W_P + W_Q; x, y; u, v) \; = \; 0$$

and the feedback map A is equal to

$$\left\{ \begin{array}{l} A(x,y;v) = \{u \in U(x,y) \mid \\ H(W_P + W_Q; x, y; u, v) = \inf_{u'\in U(x,y)} H(W_P + W_Q; x, y; u', v)\} \end{array} \right.$$

— the first player has a pure control if and only if

$$\inf_{u\in U(x,y)} \sup_{v\in V(x,y)} H(W_P + W_Q; x, y; u, v) \; = \; 0$$

and the feedback map B is equal to

$$\left\{ \begin{array}{l} B(x,y) = \{u \in U(x,y) \mid \sup_{v\in V(x,y)} H(W_P + W_Q; x, y; u, v) \\ = \inf_{u'\in U(x,y)} \sup_{v\in V(x,y)} H(W_P + W_Q; x, y; u', v)\} \end{array} \right.$$

We then deal with the main topic of this paper: construct single-valued playable feedbacks $(\tilde{u}, \tilde{v})$, such that the differential system

$$\left\{ \begin{array}{lcl} x'(t) & = & f(x(t), y(t), \tilde{u}(x(t), y(t)) \\ y'(t) & = & g(x(t), y(t), \tilde{v}(x(t), y(t)) \end{array} \right.$$

has playable solutions for each initial state. By the Playability Theorem, they must be selections of the regulation map R in the sense that

$$\forall\, (x,y) \; \in \; K, \;\; (x,y) \mapsto (\tilde{u}(x,y), \tilde{v}(x,y)) \in R(x,y)$$

We shall prove the existence of such continuous single-valued playable feedbacks, as well as more constructive, but discontinuous, playable feedbacks, such as the feedbacks associating the controls of $R(x,y)$ with minimal norm (the playable slow feedbacks, as in [13,36]). More generally, we shall show the existence of possibly set-valued feedbacks associating with any $(x,y) \in K$ the set of controls $(u,v) \in R(x,y)$ which are solutions to a (static) optimization problem of the form:

$$(u,v) \in R(x,y) \quad | \quad \sigma(x,y;u,v) \leq \inf_{u',v' \in R(x,y)} \sigma(x,y;u',v')$$

or solutions to a noncooperative game of the form:

$$\forall (u',v') \in R(x,y), \;\; a(x,u,v') \leq a(x,u,v) \leq a(x,u',v)$$

In other words, the players can implement playable solutions to the differential game by playing for each state $(x,y) \in K$ a static game on the controls of the regulation subset $R(x,y)$.

We also consider the issue of finding discriminating feedbacks, which are selections of the set-valued map A. We shall provide for instance sufficient conditions implying that for all continuous feedback $\tilde{v}(x,y) \in V(x,y)$ played by the second player, the first player can find a feedback (continuous or of minimal norm) $\tilde{u}(x,y)$ such that the above differential equation has playable solutions for each initial state.

Finally, we address the question of constructing continuous pure feedbacks $\tilde{u}(x,y)$ which have the property of yielding playable solutions of the above differential equation whatever the continuous feedback $\tilde{v}(x,y)$ played by the second player[3].

We use for constructing these feedbacks selection theorems (for instance, Michael's continuous selection theorem, see [29,30,31]), we need to prove the

[3] One can also construct "dynamic feedback controls" which are selections $(\tilde{b},\tilde{c})$ of the contingent derivative of the regulation map

$$(\tilde{b}(x,y;u,v), \tilde{c}(x,y:u,v)) \;\; \in \;\; DR(x,y)(f(x,y;u), g(x,y;v))$$

With these "dynamic feedbacks, players implement the differential system

$$\begin{cases} x'(t) & = & f(x(t),y(t),\tilde{u}(x(t),y(t)) \\ y'(t) & = & g(x(t),y(t),\tilde{v}(x(t),y(t)) \\ u'(t) & = & \tilde{b}(x(t),y(t);u(t),v(t)) \\ v'(t) & = & \tilde{c}(x(t),y(t);u(t),v(t)) \end{cases}$$

which yields playable solutions.

In other words, the players can implement playable solutions to the differential game by playing for each state $(x,y) \in K$ a static game on "velocities" of the controls in the derivative $DR(x,y)(f(x,y;u), g(x,y;v))$ of the regulation subset.

Minimal selections (b°,c°) provide heavy trajectories (see [5]) in the case of control systems lower semicontinuity of the set-valued maps R, A and B. In the case of the set-valued map B, we need a Lower Semicontinuity Criterion of an infinite intersection of lower semicontinuous maps. We provide such a theorem at the end of this paper, which can be useful for other purposes.

References

[1] AUBIN J.-P. (1987) ***Differential Calculus of set-valued maps: an update***. IIASA WP

[2] AUBIN J.-P. (1987) ***Viability Tubes and the Target Problem*** . Proceedings of the IIASA Conference, July 1986, IIASA WP-86

[3] AUBIN J.-P. & CELLINA A. (1984) DIFFERENTIAL INCLUSIONS. Springer-Verlag (Grundlehren der Math. Wissenschaften, Vol.264, 1-342)

[4] AUBIN J.-P. & EKELAND I. (1984) APPLIED NONLINEAR ANALYSIS. Wiley-Interscience

[5] AUBIN J.-P. & FRANKOWSKA H. (1985) ***Heavy viable trajectories of controlled systems***. Annales de l'Institut Heanri-Poincaré, Analyse Non Linéaire, 2, 371-395

[6] AUBIN J.-P. & FRANKOWSKA H. (1987) ***Observability of Systems under Uncertainty***. IIASA W-P-87

[7] AUBIN J.-P. & WETS R. (1986) ***Stable approximations of set-valued maps***. IIASA WP-87

[8] BERNHARD P. (1979) Contribution à l'étude des jeux différentiels à somme neulle et information parfaite. Thèse Université de Paris VI

[9] BERNHARD P. (1980) ***Exact controllability of perturbed continuous-time linear systems***. Trans. Automatic Control, 25, 89-96

[10] BLAQUIERE A. (1976) ***Dynamic Games With Coalitions and Diplomacies***. in Directions in Large-Scale Systems, 95-115

[11] CORLESS M. & LEITMANN G. (1984) ***Adaptive Control for Uncertain Dynamical Systems***. Dynamical Systems and Microphysics Control Theory and Mechanics, 91-158

[12] CORLESS M. , LEITMANN G. & RYAN E.P. (1984) ***Tracking in the Presence of Bounded Uncertainties***. Proceeding of the Fourth International Conference on Control Theory

[13] FALCONE M. & SAINT-PIERRE P. (1987) *Slow and quasi-slow solutions of differential inclusions.* J. Nonlinear Anal.,T.,M.,A., 3, 367-377

[14] FRANKOWSKA H. (1987) *L'équation d'Hamilton-Jacobi contingente.* Comptes Rendus de l'Académie des Sciences, PARIS,

[15] FRANKOWSKA H. (1987) *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi.* IIASA WP-87-069

[16] GUSEINOV H. G. , SUBBOTIN A. I. & USHAKOV V. N. (1985) *Derivatives for multivalued mappings with applications to game theoretical problems of control.* Problems of Control and Information Theory, Vol.14, 155-167

[17] HADDAD G. (1981) *Monotone trajectories of differential inclusions with memory.* Israel J. Maths, 39, 38-100

[18] ISSACS R. (1965) DIFFERENTIAL GAMES. Wiley, New York

[19] KRASOVSKI N. N. (1986) THE CONTROL OF A DYNAMIC SYSTEM. Nauka, Moscow

[20] KRASOVSKI N. N. & SUBBOTIN A. I. (1974) POSITIONAL DIFFERENTIAL GAMES. Nauka, Moscow

[21] KURZHANSKII A. B. (1977) CONTROL AND OBSERVATION UNDER CONDITIONS OF UNCERTAINTY. Nauka

[22] KURZHANSKII A. B. (1986) *On the analytical properties of viability tubes of trajectories of differential systems.* Doklady Acad. Nauk SSSR, 287, 1047-1050

[23] KURZHANSKII A. B. & FILIPPOVA T. F. (1986) *On viable solutions for uncertain systems.*

[24] LEDYAEV YU.S. (1985) *Regular differential games with mixed constraints on the controls.* Proceedings of the Steklov Institute of Mathematics, 167, 233-242

[25] LEITMANN G. (1980) *Guaranteed avoidance strategies.* Journal of Optimization Theory and Applications, Vol.32, 569-576

[26] LEITMANN G. (1981) THE CALCULUS OF VARIATIONS AND OPTIMAL CONTROL. Plenum Press

[27] LEITMANN G. , RYAN E.P. & STEINBERG A. (1986) *Feedback control of uncertain systems: robustness with respect to neglected actúator and sensor dynamics.* Int. J. Control, Vol.43, 1243-1256

[28] LIONS P. -L. (1982) GENERALIZED SOLUTIONS OF HAMILTON-JACOBI EQUATIONS. Pitman

[29] MICHAEL E. (1956) *Continuous selections I.* Annals of Math., 63, 361-381

[30] MICHAEL E. (1956) *Continuous selections II.* Annals of Math., 64, 562-580

[31] MICHAEL E. (1957) *Continuous selections III.* Annals of Math., 65, 375-390

[32] RAY A. & BLAQUIERE A. (1981) *Sufficient Conditions for Optimality of Threat Strategies in a Differential Game.* Journal of Optimization Theory and Applications, 33, 99-109

[33] SUBBOTIN A. I. (1985) *Conditions for optimality of a guaranteed outcome in game problems of control.* Proceedings of the Steklov Institute of Mathematics, 167, 291-304

[34] SUBBOTIN A. I. & SUBBOTINA N. N. (1983) *Differentiability properties of the value function of a differential game with integral terminal costs.* Problems of Control and Information Theory, 12, 153-166

[35] SUBBOTIN A. I. & TARASYEV A. M. (1986) *Stability properties of the value function of a differential game and viscosity solutions of Hamiltopn-Jacobi equations.* Problems of Control and Information Theory, 15, 451-463

[36] TCHOU N. A. (1988) *Existence of slow monotone solutions to a differential inclusion.* J. Math. Anal. Appl.

DYNAMIC OPTIMIZATION OF SOME FORWARD-LOOKING STOCHASTIC MODELS[†]

Tamer Başar

Decision and Control Laboratory
Coordinated Science Laboratory and the
Department of Electrical and Computer Engineering
University of Illinois
1101 W. Springfield Avenue
Urbana, Illinois 61801 / USA

Abstract

A dynamic decision model is said to be *forward-looking* if the evolution of the underlying system depends explicitly on the expectations the agents form on the future evolution itself. Such models lead to nonstandard stochastic dynamic optimization problems where one has to take into account the fact that there is a circular (closed) relationship between future forecasts and future system behavior. In this paper we study a class of such problems where there is an additional control input designed to make the system track a given trajectory. This leads to a game-theoretic formulation in which context we consider both finite and infinite horizon formulations. It is shown that for the finite horizon problem the unique Nash equilibrium solution requires (fixed size) memory for both agents because of spillover across stages, whereas for the infinite horizon version no memory is needed.

1. An Introduction to Forward-Looking Models

We refer to a dynamic stochastic model as *forward-looking* if one of its inputs involves future expectations of the system trajectory, using (possibly noisy) measurements on the past realizations. Such decision models find wide-spread use in economics, where they are more commonly known as *rational expertations* models. A few representative papers in this area are the works of *Lucas (1975)*, *Sargent and Wallace (1975)*, *Barro (1976)*, *Taylor (1977)*, *Shiller (1978)*, *Blanchard (1979)*, and *Blanchard and Kahn (1980)*. Perhaps the simplest such model that captures the salient features of *forward-looking* behavior is described by the scalar difference equation

$$y_{t+1} = ay_t + bv_t + \epsilon_{t+1}, \tag{1a}$$

where a and b are constant parameters, $\{\epsilon_t\}$ is a sequence of independent zero-mean random variables with finite (positive) variance, and v_t is the decision variable chosen at time t under some "expectation" of the future behavior of the system based on information available at time t. If the forecast of interest is n steps into the future, for example, one possibility is to replace v_t in $(1a)$ by $E_t y_{t+n}$, the conditional mean of y_{t+n} based on the information available at time t. This information, which we denote by η_t, could involve a direct measurement of all the past values of the system trajectory, that is $\{y_t, y_{t-1}, ...\} =: y^t$, or involve some *noisy* measurement on the state trajectory, $\eta_t = z^t$, where $\{z_t\}$ is a measurement sequence defined by

[†] *This work was performed while the author was spending a sabbatical year at INRIA, Sophia Antipolis, France, and it was also partially supported by the Air Force Office of Scientific Research under Grant No. AFOSR 084-0056, through the University of Illinois.*

$$z_t = y_t + \xi_t, \tag{1b}$$

with $\{\xi_t\}$ being another sequence of independent, zero-mean random variables with finite variance.

A basic question addressed in the literature over the years has been the existence of a (unique) stochastic process $\{y_t\}$ that satisfies (1*a*) whenever $v_t = E_t y_{t+n}$ and the time interval is infinite. The answer to this question is that there is, in general, more than one such solution even in the class of *stationary* processes. However, as we have recently argued in *Başar (1987)*, a better approach towards policy determination in these *forward-looking* models would involve the optimization of an appropriate loss function, by carefully taking into account the informational dependence as well as the correlation of policies across stages. One such criterion would be

$$J_s^T := \min_{\{\gamma_t\}} \sum_{t=s}^{T} E\{[\gamma_t(\eta_t) - y_{t+n}]^2\}\rho^{t-s}, \tag{2}$$

where minimization is subject to the dynamic constraint (1*a*), with $v_t = \gamma_t(\eta_t)$, and uses the boundary condition $v_t \equiv 0$ for $t > T$. In the above, $[s, T]$ stands for the time horizon, which could also be infinite, and ρ denotes a positive discount factor $(0 < \rho \leq 1)$. It has been shown in *Başar (1987)* that the dynamic policy optimization problem admits the solution $v_t = E_t y_{t+1}$ when $n = 1$, but for $n \geq 2$ the unique solution for the finite-horizon version is different from $E_t y_{t+n}$. For $n = 2$, for example, the best forecast into the future (by *two* time steps), under the criterion (2) and using the information $\{\eta_t = y^t\}$, is given by

$$v_t^* = \gamma_t^*(y^t) = \alpha_t y_t + \beta_t v_{t-1}, \tag{3a}$$

for $2 \leq t \leq T$, where the sequences $\{\alpha_t\}$ and $\{\beta_t\}$ are determined recursively *off-line*. For the *noisy* measurement case, $\{\eta_t = z^t\}$, the solution is again unique and is given by

$$v_t^* = \gamma_t^*(z^t) = \alpha_t \hat{y}_t + \beta_t v_{t-1}, \tag{3b}$$

for $2 \leq t \leq T$, where the sequences $\{\alpha_t\}$ and $\{\beta_t\}$ are the same as in (3*a*), and $\hat{y}_t$ is a sequence of estimates generated recursively by a Kalman filter, under the assumption that the underlying statistics are Gaussian. An interesting feature of the solution is that for the infinite-horizon version (that is as $T \to \infty$) the coefficient sequence $\{\beta_t\}$ vanishes for all finite t, and the solution becomes $v_t^* = E_t y_{t+2}$,thus eliminating the correlation across stages.

In the present paper, we consider a more general formulation than that above, where now two separate agents, say A and B, have influence on the system trajectory, one of them (A) again making a two-step ahead forecast of the trajectory, whereas the other one (B) trying to drive the trajectory as close to a specific target as possible. For such a scenario, the system equation would be replaced by

$$y_{t+1} = a y_t + b v_t + c w_t + \epsilon_{t+1}, \tag{4}$$

where $v_t = \gamma_t(y^t)$ is controlled by agent A and $w_t = \mu_t(y^t)$ by agent B. Taking the time horizon as $[0, T+1]$, the two cost functions to be minimized by A and B, respectively, are

$$J_A(\gamma, \mu) = \sum_{t=0}^{T} E\{[\gamma_t(y^t) - y_{t+2}]^2\}\rho_A^t, \tag{5a}$$

and

$$J_B(\gamma,\mu) = \sum_{t=0}^{T+1} E\{[y_{t+1} - \bar{y}_{t+1}]^2 + kw_t^2\}\rho_B^t, \tag{5b}$$

where $\{\bar{y}_t, 2 \leq t \leq T+2\}$ is the desired trajectory, k is a positive weight on agent B's control, ρ_A, ρ_B are the corresponding discount factors, $\gamma := \{\gamma_T, \gamma_{T-1}, .., \gamma_0\}$, $\mu := \{\mu_{T+1}, \mu_T, .., \mu_0\}$, and $v_{T+1} \equiv 0$, the last identity reflecting the fact that no forecast is made at time $t = T+1$. Furthermore, we assume that the independent random variables ϵ_t $(1 \leq t \leq T+2)$ each have zero mean and a probability distribution that assigns positive probability to every measurable open subset of the real line. One such distribution would be the normal (Gaussian) distribution with positive variance.

Since this is a problem with multiple objectives, several equilibrium solution concepts would be applicable, with the one adopted here being the noncooperative *Nash* equilibrium solution. Therefore, we seek a pair (γ^*, μ^*), preferably unique, satisfying the pair of inequalities

$$J_A(\gamma^*,\mu^*) \leq J_A(\gamma,\mu^*); \quad J_B(\gamma^*,\mu^*) \leq J_B(\gamma^*,\mu), \tag{6}$$

for all admissible γ and μ. Other possibilities would have been the Stackelberg solution with either agent acting as the leader and the Pareto-optimal solution, which, however, we do not discuss here because of space limitations.

The first question we attack, in *section 2*, is a "simpler" version of the above, where agent A's policy is fixed as $v_t = E_t y_{t+2}, t \leq T$, which is in general a suboptimal policy for A. We obtain the best policy for B under this additional structural restriction, and derive the corresponding expression for $\{\mu_t\}$ (see *Theorem 1*). Furthermore, we study the limiting behavior of the two policies, for the infinite-horizon problem. Subsequently, in *section 3*, we derive the unrestricted Nash solution and prove its (generic) existence and uniqueness (see *Theorem 2*), with details of the derivation provided in the *Appendix*. We also study the limiting behavior of the solution as $T \to \infty$, and analyze the discrepancies that exist between the two stationary solutions of *Theorem 1* and *Theorem 2*. The paper concludes with a discussion of the "noisy measurement" case and some other possible extensions, in *section 4*. Throughout the analysis, we take the reference trajectory (to be tracked) as the *zero* trajectory, an assumption that does not bring in much loss of conceptual generality but leads to considerable simplifications in the resulting expressions.

2. The Optimal Tracking Strategy Under Perfect Myopic Forecast

With v_t taken as $E_t y_{t+2}$ (*which myopically minimizes each term of* (*5a*)), and $\{\bar{y}_t\}$ taken as the *zero* trajectory, the dynamic policy optimization problem faced by agent B is the minimization of F_0^{T+1}, where

$$F_s^{T+1} = \sum_{t=s}^{T+1} E\{y_{t+1}^2 + kw_t^2\}\rho_B^{t-s}, \tag{7a}$$

the dynamic constraint is

$$\begin{aligned} y_{T+2} &= ay_{T+1} + cw_{T+1} + \epsilon_{T+2}, \\ y_{t+1} &= ay_t + bE_t y_{t+2} + cw_t + \epsilon_{t+1}, \quad t \leq T, \end{aligned} \tag{7b}$$

and the information constraint is $w_t = \mu_t(y^t)$. Note that this is not a standard linear-quadratic stochastic control problem because of the presence of the conditional expectations term in (7*b*), which could even make the dynamic constraint nonlinear in the past values of the trajectory. We will show below, however, that the optimal control is still linear, thus making the corresponding forecast also linear in the available information. The derivation entails a recursive approach where the structure of v_t is determined alongside the optimal control at each step of the iterative minimization.

Before presenting the main result of this section, we first introduce two sequences $\{p_t\}$ and $\{\nu_t\}$ which are defined recursively by

$$p_t = 1 + \frac{\rho_B a^2 k p_{t+1}}{c^2 p_{t+1} + k\nu_t^2}; \quad p_{T+2} = 1, \tag{8a}$$

$$\nu_{t-1} = 1 - \frac{abk\nu_t}{c^2 p_{t+1} + k\nu_t^2}; \quad \nu_{T+1} = 1. \tag{8b}$$

Next we define a third sequence $\{g_t\}$ in terms of the other two, according to

$$g_t = -cap_{t+1}/(c^2 p_{t+1} + k\nu_t^2), \quad t \leq T+1. \tag{8c}$$

We are now in a position to state the main result, after invoking a condition which generically holds.

Condition 1. *The sequence $\{\nu_t\}$ generated by (8b) does not vanish for any $t \leq T+1$.*

Theorem 1. *Let Condition 1 be satisfied. Then, the dynamic policy optimization problem with myopic forecast admits the unique solution*

$$w_t = \hat{\mu}_t(y_t) = g_t y_t, \quad 0 \leq t \leq T+1, \tag{9a}$$

with the corresponding forecast policy given by

$$v_t = E_t y_{t+2} = \frac{(a + cg_{t+1})(a + cg_t)}{\nu_{t+1}\nu_t} y_t := h_t y_t. \tag{9b}$$

The minimum value of F_0^{T+1} in (6a) is

$$\hat{F}_0^{T+1} = p_0 E\{y_0^2\} + \lambda_0, \tag{9c}$$

where λ_0 is the last step of the backward recursion

$$\lambda_{T+1} = var(\epsilon_{T+2}),$$
$$\lambda_{t-1} = \rho_B \lambda_t + p_t var(\epsilon_t).$$

Proof. The proof proceeds by recursively showing that the minimum value of F_s^{T+1} is given, for each $s \leq T+1$, by

$$\hat{F}_s^{T+1} = [(p_s - 1)/\rho_B]E\{y_s^2\} + \lambda_s.$$

The result is trivially true for $s = T+2$, where we take $\lambda_{T+2} = 0$. Let us therefore assume its validity, along with (9a) and (9b), up to $s+1$, and verify the expression, as well as (9a) and (9b), for s. The minimization problem faced by agent B at time s is

$$\min_{\mu_s}[\rho_B \hat{F}_{s+1}^{T+1} + E\{y_{s+1}^2 + k\mu_s(y^s)^2\}], \tag{*}$$

which is equivalent to

$$\min_{w} E\{p_{s+1}y_{s+1}^2 + kw^2 \mid y^s\},$$

which uses the dynamic constraint

$$y_{s+1} = ay_s + bE_s y_{s+2} + cw + \epsilon_{s+1}. \qquad (\star)$$

We also have the relationship

$$y_{s+2} = (a + cg_{s+1})y_{s+1} + b[(a + cg_{s+2})(a + cg_{s+1})/\nu_{s+2}\nu_{s+1}]y_{s+1} + \epsilon_{s+2},$$

where we have explicitly used (9a) and (9b), with t replaced by $s+1$. (Of course, if $s = T+1$, the last relationship would not be needed since the conditional expectation term in ($\star$) would be missing.) Now, taking the conditional expectation of the last expression with respect to y^s, substituting this into ($\star$), taking the conditional expectation of the resulting expression again with respect to y^s, and solving for the resulting $E_s y_{s+1}$ in terms of y_s and w, we arrive at the expression

$$E_s y_{s+1} = \frac{1}{\nu_s}[ay_s + cw].$$

Using this, $E_s y_{s+2}$ can easily be evaluated to be

$$E_s y_{s+2} = \frac{a + cg_{s+1} + bh_{s+1}}{\nu_s}[ay_s + cw], \qquad (\star\star)$$

under which the dynamic constraint becomes equivalent to

$$y_{s+1} = \frac{1}{\nu_s}[ay_s + cw] + \epsilon_{s+1}.$$

This makes the minimization problem a standard *linear-quadratic* one, and hence it readily follows that the minimizing control is uniquely given by (9a) with $t = s$. Substitution of this solution into (•) and ($\star\star$) finally verifies the asserted form for $\hat{F}_s^{T+1}$ and the structure of the forecast policy as given by (9b). We should note that *Condition 1* has explicitly been used in the proof, to make sure that one can solve uniquely for $E_s y_{s+1}$ and $E_s y_{s+2}$. ◇

Condition 1, under which the existence and uniqueness of the optimal control (9a) is valid, holds whenever a and b have opposite signs, regardless of the magnitudes of the parameters of the problem. The result follows by inspection, since with $ab < 0$ and $\nu_{T+1} = 1$, we have $\nu_t > 0$ for all $t \leq T+1$. For $ab < 0$, however, there may exist isolated values for the parameters for which the condition does not hold for some t. [A more precise statement here would be that with all but one of the parameters fixed (and $ab > 0$), there will exist at most a finite number of different values of that parameter for which *Condition 1* is violated. This follows since for each t, ν_t is a rational function of the quintuplet (a, b, c, ρ_B, k).] For example, for the parameter values $a = c = k = 1$, $b = 2$, we have $\nu_T = 0$, which shows that *Condition 1* may fail even for a two-stage problem. However, if we perturb the value of b to $b = 2.1$, and take $\rho_B = 1$, then *Condition 1* holds for all values of t. In fact, running the coupled recursive equations (8a)-(8b) in retrograde time, we find that (for these parameter values) the pair (ν_t, p_t) converges to $(0.504147, 1.880960)$ in 29 steps, within the accuracy of six decimal places. Hence, in this case, the infinite-horizon version (even with no discounting) admits a unique optimal stationary control, given by $w_t = \hat{\mu}(y_t) = -1.135124\, y_t$. If, in the above, b is instead taken to be 1, again *Condition 1* holds, the pair (ν_t, p_t) converges to $(0.694146, 1.787692)$ in 9 iterations, and the optimal control policy converges to $w_t = \hat{\mu}(y_t) = -0.787692\ y_t$. As a final numerical experimentation, reflecting a different set of parameter values, we consider the case of $a = 2, b = -3, \rho_B = 0.8, c = k = 1$. For this set, we already know that *Condition 1* holds, since $ab < 0$. Studying the convergence of the optimal policy to a stationary control, we find that the pair (ν_t, p_t) converges to $(2.796267, 1.521150)$ in 26 iterations, with the resulting stationary policy being $w_t = \hat{\mu}(y_t) = -0.325719\ y_t$.

3. The Nash Equilibrium Solution

We now remove the restriction that agent A's input to the system is a myopic forecast, and allow him to determine the "best" choice for v_t by minimizing the cost function J_A. As we have discussed in *section 1*, this joint optimization problem can best be treated as a noncooperative game, and hence we study in this section the Nash equilibrium of the underlying game, as defined by (6).

There are two general approaches to the derivation of Nash equilibria in such dynamic games. One would be first to guess (or propose) a structure for the solution in terms of some parameters, and then to validate the equilibrium property of the asserted structure and to obtain the corresponding values of the parameters so that the resulting policies are in Nash equilibrium. A second approach would be to obtain the Nash solution recursively (by employing the definition of *stagewise* or *feedback* equilibrium; see, for example, *Başar and Olsder (1982)*) by solving static games conditioned on the available (common) information, at each step of the iteration. Note that this would be applicable only if both agents have identical information (which is the case here), since otherwise stagewise decomposition would not be possible. Now, two disadvantages of the first method are that (i) one has to guess the structure of the solution correctly, and (ii) even if the initial guess is correct there is no way to show (using this method) that the validated Nash solution is unique. The second method, on the other hand, is capable of answering the uniqueness question, but it only produces candidate solutions which subsequently have to be checked for their equilibrium property. What we will, therefore, choose to do in the sequel is to use an appropriate combination of the two approaches, to generate candidate solutions and verify their existence and uniqueness. We should note in passing that even though the problem may look, at the outset, as a standard *linear-quadratic* one, the presence of the two-step delay in the cost function of agent A makes the game a nonstandard one, thus eliminating the possibility of direct application of results available on linear-quadratic feedback Nash games (as, for example, covered in *Başar and Olsder (1982)*).

Before presenting the solution in *Theorem 2* below, we first introduce some sequences which will be needed in the characterization of the equilibrium policies. Towards this end, let $\{m_t\}$, $\{\tilde{m}_t\}$, $\{n_t\}$ be three sequences generated by

$$m_{t-1} = \frac{1}{1 - bm_t}[a + c\tilde{\alpha}_t + b\alpha_t\tilde{m}_t]; \quad m_T = \frac{ak}{k + c^2}, \tag{10a}$$

$$\tilde{m}_{t-1} = \frac{1}{1 - bm_t}[c\tilde{\beta}_t + b\beta_t\tilde{m}_t]; \quad \tilde{m}_T = 0, \tag{10b}$$

$$n_{t-1} = \frac{\rho_A n_t (1 - bm_t)^2}{b^2 + \rho_A n_t (1 - bm_t)^2}; \quad n_T = 1, \tag{10c}$$

where α_t, $\tilde{\alpha}_t$, β_t, $\tilde{\beta}_t$ are defined, for $t \leq T$, by

$$\alpha_t := \frac{(1 - bm_t)\rho_A n_t m_t - b}{b^2 + \rho_A n_t (1 - bm_t)(1 - bm_t - \tilde{m}_t)}[a + c\tilde{\alpha}_t], \quad t \leq T, \tag{11a}$$

$$\widetilde{\alpha}_{t-1} := -\frac{[ak_{11,t} + (bk_{11,t} + k_{22,t})\alpha_{t-1}]c}{k + c^2 k_{11,t}}, \qquad t \le T+1, \tag{11b}$$

$$\beta_t := \frac{(1 - bm_t)\rho_A n_t m_t c\widetilde{\beta}_t + b(1 - c\widetilde{\beta}_t)}{b^2 + \rho_A n_t (1 - bm_t)(1 - bm_t - \widetilde{m}_t)}, \quad t \le T, \tag{12a}$$

$$\widetilde{\beta}_{t-1} := -\frac{(bk_{11,t} + k_{12,t})c\beta_{t-1}}{k + c^2 k_{11,t}}, \qquad t \le T+1, \tag{12b}$$

and

$$K_t := \begin{pmatrix} k_{11,t} & k_{12,t} \\ k_{12,t} & k_{22,t} \end{pmatrix}$$

is a 2 × 2 matrix sequence generated by the discrete time Riccati equation

$$\begin{aligned} K_t &= \rho_B A_t' [K_{t+1} - K_{t+1} C (C' K_{t+1} C + k)^{-1} C' Kt + 1] A_t + Q, \\ K_{T+1} &= [k a^2 \rho_B / (k + c^2)] Q, \end{aligned} \tag{13a}$$

with

$$A_t := \begin{pmatrix} a + b\alpha_t & b\beta_t \\ \alpha_t & \beta_t \end{pmatrix}, \quad Q := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad C := \begin{pmatrix} c \\ 0 \end{pmatrix}. \tag{13b}$$

Finally, let $r_{\alpha t}$, $r_{\tilde{\alpha} t}$, $r_{\beta t}$, $r_{\tilde{\beta}_t}$ be defined by

$$r_{\alpha t} := \frac{c[(1 - bm_t)\rho_A m_t n_t - b]}{b^2 + \rho_A n_t (1 - bm_t)(1 - bm_t - \widetilde{m}_t)}, \tag{14a}$$

$$r_{\tilde{\alpha} t} := -\frac{c(bk_{11,t+1} + k_{22,t+1})}{k + c^2 k_{11,t+1}}, \tag{14b}$$

$$r_{\beta t} := \frac{c(1 - bm_t)\rho_A n_t m_t - bc}{b^2 + \rho_A n_t (1 - bm_t)(1 - bm_t - \widetilde{m}_t)}, \tag{15a}$$

$$r_{\tilde{\beta} t} := -\frac{c(bk_{11,t+1} + k_{12,t+1})}{k + c^2 k_{11,t+1}}. \tag{15b}$$

The last four expressions are the coefficient terms in (11*a*)-(12*b*), indicating the dependence of α_t, $\widetilde{\alpha}_t$, β_t and $\widetilde{\beta}_t$ on $\widetilde{\alpha}_t$, α_t, $\widetilde{\beta}_t$ and β_t, respectively. A certain relationship between these coefficient terms in fact determines the existence of a unique Nash equilibrium solution, as to be elucidated below.

Condition 2. *For all $t \le T$,*

$$r_{\alpha t} r_{\tilde{\alpha} t} \neq 1, \quad r_{\beta t} r_{\tilde{\beta}_t} \neq 1, \tag{16a}$$

$$bm_t \neq 1, \tag{16b}$$

$$\rho_A n_t (1 - bm_t)(1 - bm_t - \widetilde{m}_t) \neq -b^2. \tag{16c}$$

Theorem 2. *Let Condition 2 be satisfied. Then, the forward-looking tracking model* (4)-(5) *admits a unique Nash equilibrium solution $\{\gamma_t^*, \mu_t^*\}$, where agent A's (best forecast) policy is*

$$\begin{aligned} v_t = \gamma_t^*(y^t) &= \alpha_t y_t + \beta_t \widetilde{v}_{t-1}, \quad && t \ge 1 \\ &= \alpha_0 y_0, && t = 0, \end{aligned} \tag{17}$$

and agent **B***'s (best tracking) policy is*

$$\begin{aligned} w_t = \mu_t^*(y^t) &= \widetilde{\alpha}_t y_t + \widetilde{\beta}_t \widetilde{v}_{t-1}, && 1 \le t \le T \\ &= -[ac/(k+c^2)]y_{T+1}, && t = T+1 \\ &= \widetilde{\alpha}_0 y_0, && t = 0 \quad , \end{aligned} \tag{18a}$$

where the sequence $\{\widetilde{v}_t\}$ *is generated by*

$$\begin{aligned} \widetilde{v}_t = &= \alpha_t y_t + \beta_t \widetilde{v}_{t-1}, && t \ge 1 \\ &= \alpha_0 y_0, && t = 0. \end{aligned} \tag{18b}$$

Proof. We will first verify the structural consistency of the solution (17)-(18) under the Nash inequalities (6), and then discuss the existence of the solution. Some details of the derivation, as well as a proof for the uniqueness of the solution will be given in the *Appendix*.

Towards verifying the validity of (6), first consider the second inequality, where agent A's policy has been fixed as given by (17). Then, agent B faces a stochastic control problem with cost function J_B (given by (5b) with *zero* reference trajectory) and state dynamics

$$\begin{aligned} y_{t+1} &= (a + b\alpha_t)y_t + b\beta_t \widetilde{v}_{t-1} + cw_t + \epsilon_{t+1} \quad , && t \le T \\ &= ay_{T+1} + cw_{T+1} + \epsilon_{T+2} \quad , && t = T+1, \end{aligned}$$

where the sequence $\{\widetilde{v}_t\}$ is generated by (18b) in view of (17). The optimal control at time $T+1$, w_{T+1}, can readily be obtained, to be given by the second line in (18a). To obtain the remaining controls, we introduce a new state vector, $x_t := (y_t, \widetilde{v}_{t-1})$, and reformulate the problem as one of minimizing J_B under the dynamic constraint

$$x_{t+1} = A_t x_t + Cw_t + D\epsilon_{t+1} \quad ; \qquad D := (1 \quad 0)',$$

where control w_t is allowed to depend on x^t, $t \le T$. [Note that even though v^{t-1} is not available to agent B, $\widetilde{v}^{t-1}$ is since it is generated by y^{t-1}.] This is the familiar linear-quadratic optimal control problem, whose unique solution is

$$w_t = -(k + C'K_{t+1}C)^{-1}C'K_{t+1}A_t x_t, \tag{$\star$}$$

where $\{K_t\}$ is generated by (13a). [Note that the terminal constraint on K_t at $t = T+1$ is not Q because we have already substituted for the optimal w_{T+1} and have reduced the cost function J_B to the one where the leading term is now y_{T+1}^2 instead of y_{T+2}^2.] Now, the optimal control ($\star$) is clearly linear in y_t and $\widetilde{v}_{t-1}$, at time t, and a little algebra shows that it can be expressed in the form (18a).

We now focus attention on the first inequality of (6), where agent B's policy is fixed as given by (18a). Then the problem faced by agent A is one of optimal forecast, where the cost function is J_A (given by (5a)) and the dynamic constraint is

$$\begin{aligned} y_{t+1} &= (a + c\widetilde{\alpha}_t)y_t + c\widetilde{\beta}_t \widetilde{v}_{t-1} + bv_t + \epsilon_{t+1}, && 1 \le t \le T \\ &= -[ak/(k+c^2)]y_{T+1} + \epsilon_{T+2}, && t = T+1 \\ &= (a + c\widetilde{\alpha}_0)y_0 + bv_0 + \epsilon_1, && t = 0. \end{aligned}$$

Because of the form of the cost function, the available linear-quadratic theory cannot be directly applied to this problem; nevertheless, a one can employ a *dynamic programming* type argument to construct the optimal solution in retrograde time, as in the proof of *Theorem 2.1* of *Başar (1987)*. It has been shown in the *Appendix* that the optimal solution is unique (under some conditions which will be specified later), and the optimal policy at time t is a function of three variables, y_t, $\widetilde{v}_{t-1}$ and v_{t-1}. The precise expression is

$$\begin{aligned} v_t = \gamma_t(y^t) &= \widehat{\alpha}_t y_t + \widehat{\beta}_t v_{t-1} + \bar{\beta}_t \widetilde{v}_{t-1}, \quad 1 \le t \le T \\ &= \widehat{\alpha}_0 y_0, \qquad\qquad\qquad\qquad t = 0, \end{aligned} \tag{$**$}$$

where

$$\widehat{\alpha}_t = \frac{1}{b^2 + (1 - bm_t)^2 \rho_A n_t} [\rho_A n_t (1 - bm_t)(\widetilde{m}_t \alpha_t + (a + c\widetilde{\alpha}_t) m_t) - b(a + c\widetilde{\alpha}_t)] \tag{$\bullet$}$$

$$\widehat{\beta}_t = \frac{b}{b^2 + (1 - bm_t)^2 \rho_A n_t}$$

$$\bar{\beta}_t = \frac{1}{b^2 + (1 - bm_t)^2 \rho_A n_t} [\rho_A n_t (1 - bm_t)(\widetilde{m}_t \beta + c\widetilde{\beta}_t m_t) - bc\widetilde{\beta}_t], \tag{$\bullet\bullet$}$$

and $\{m_t\}$, $\{\widetilde{m}_t\}$, $\{n_t\}$ are generated by (10*a*)-(10*c*). In writing down these expressions, we have already assumed the validity of (16*a*) and (16*c*), since otherwise m_t and $\widetilde{m}_t$ would not have been well defined. We should note, however, that even in the *pure* forecast problem discussed in *Başar (1987)*, a condition similar to (16*b*) was required for the well-posedness of the problem.

Now, to complete the derivation, we substitute for α_t and β_t in ($\bullet$) and ($\bullet\bullet$) from (11*a*) and (12*a*), respectively, and observe that the resulting expression for $\widehat{\alpha}_t$ is identical with that of α_t, and also when the resulting expression for $\bar{\beta}_t$ is added to $\widehat{\beta}_t$, the outcome is identical to β_t; in other words,

$$\widehat{\alpha}_t \equiv \alpha_t \quad , \quad \widehat{\beta}_t + \bar{\beta} \equiv \beta_t .$$

When the latter is used in ($**$) recursively, it follows that $\{v_t\}$ is generated by the same sequence (of y_s's) as $\{\widetilde{v}_t\}$, and hence that ($**$) admits the simpler representation (17).

This then completes the verification of the existence part of the theorem; more precisely, of the fact that the policies (17)-(18) constitute a Nash equilibrium pair under *Condition 2*. Note that (16*a*) in *Condition 2* simply guarantees that there is a unique solution to the two pairs of coupled equations (11) and (12), for all t, and it may also be referred to as the *Nash condition*.

As we have indicated earlier, the uniqueness part of the theorem has been verified separately in the *Appendix*. ◇

Several observations and remarks would be in order here. Firstly, we note that, as opposed to the *memoryless* solution of *Theorem 1* (obtained under myopic forecast), the unique Nash equilibrium solution incorporates memory, for both agents. For agent A, the "best" forecast policy is a linear function of the most recent measurement and the most recent decision taken by that agent. [This is true since $\widetilde{v}_{t-1}$ in (17) can be replaced by v_{t-1}, without affecting the solution.] For agent B, on the other hand, the "best" tracking policy is a linear function of the most recent measurement and a linear aggregate of all past measurements, weighted in an appropriate manner. The solution is characterised in terms of four gain coefficients $(\alpha_t, \widetilde{\alpha}_t, \beta_t, \widetilde{\beta}_t)$, which can be computed recursively. Hence, the solution does not change structurally over time, which makes it feasible to obtain *stationary* Nash policies for the infinite-horizon version, provided that the sequences $\{\alpha_t^T\}$, $\{\widetilde{\alpha}_t^T\}$, $\{\beta_t^T\}$, $\{\widetilde{\beta}_t^T\}$ converge for all finite t as $T \to \infty$, where the superscript T in the sequences denotes the dependence of each sequence on the terminal time, taken as a parameter. Even though the computation of the four critical quantities $(\alpha_t, \tilde{\alpha}_t, \beta_t, \tilde{\beta}_t)$ may look complicated at the outset, the iterations are in fact quite straightforward, requiring simple algebraic manipulations at each step. The order one has to follow in the computation is as follows:

> *Starting at t=T, first compute the quadruple $(\alpha_T, \tilde{\alpha}_T, \beta_T, \tilde{\beta}_T)$ from (11a)-(12b), using the given boundary conditions on K_{T+1}, m_T, $\tilde{m}_T$ and n_T. Note that this computation involves the solution of two pairs of coupled linear equations, at which point we invoke the Nash condition (16b) to obtain a unique solution. At this stage also condition (16c) is invoked, so that (11a) and (12a) are well defined. The next step would be to obtain the new values for k_{t+1}, m_t, $\tilde{m}_t$, n_t at t=$T+1$, using the iterations (13a), (10a), (10b) and (10c), respectively. At this stage, condition (16a) is invoked so that (10a) and (10b) are well-defined relationships. These new values for K, m, $\tilde{m}$, n are then used again in (11a)-(12b) to update the values of the gain coefficients, and this procedure is repeated until the initial stage $t = 0$ is reached.*

We should point out that similar to *Condition 1* in *Section 2*, *Condition 2* also generically holds, in the sense that if all but one of the parameter values are fixed, then there is only a finite number of values for that parameter for which the condition fails.

Even though it is not our intention to provide here a general convergence analysis for the infinite-horizon problem (this would in fact be quite a challenging task), it would nevertheless be instructive to study some properties of the stationary solution, assuming that such a solution exists and *Condition 2* holds for all t of interest. Accordingly, letting

$$\alpha^* := \lim_{T\to\infty} \alpha_t^T, \quad \tilde{\alpha}^* := \lim_{T\to\infty} \tilde{\alpha}_t^T, \quad \beta^* := \lim_{T\to\infty} \beta_t^T, \quad \tilde{\beta}^* := \lim_{T\to\infty} \tilde{\beta}_t^T, \quad n^* := \lim_{T\to\infty} n_t^T,$$

it readily follows that n^*=0. In view of this, we arrive at the stationary Nash policies

$$v_t = \gamma^*(y^t) = \alpha^* y_t + \beta^* v_{t-1}, \tag{19a}$$

$$w_t = \mu^*(y^t) = \tilde{\alpha}^* y_t + \tilde{\beta}^* \tilde{v}_{t-1}, \tag{19b}$$

where $\{\tilde{v}_t\}$ is generated by

$$\tilde{v}_t = \alpha^* y_t + \beta^* \tilde{v}_{t-1}, \tag{19c}$$

and the following relationship holds:

$$\alpha^* = -\frac{a + c\tilde{\alpha}^*}{b} \quad , \quad \beta^* = \frac{1 - c\tilde{\beta}^*}{b} \quad . \tag{20}$$

Now, using these stationary policies in the system equation (4), we arrive at the result that the equilibrium trajectory $\{y_t^*\}$ is generated by

$$y_{t+1}^* = (a + b\alpha^* + c\tilde{\alpha}^*)y_t + b\beta^* v_{t-1}^* + c\tilde{\beta}^* \tilde{v}_{t-1}^* + \epsilon_{t+1},$$

where $\{v_t^*\}$ and $\{\tilde{v}_t^*\}$ denote the discrete-time stochastic processes generated by (19a) and (19b), respectively, when $y_t = y_t^*$, $t \geq 0$. Note that, as stochastic processes, they are identical *almost surely*, and hence, by also using (20), it can be shown that the equilibrium trajectory $\{y_t^*\}$ is generated by the simpler dynamics

$$\begin{aligned} y_{t+1}^* &= v_{t-1}^* + \epsilon_t \\ v_t^* &= \alpha^* y_t^* + \beta^* v_{t-1}^* \quad , \end{aligned}$$

which admits the *ARMA* representation

$$y^*_{t+1} + \beta^* y^*_t - \alpha^* y^*_{t-1} = \epsilon_{t+1} + \beta^* \epsilon_t. \tag{21}$$

An important observation that can be made here is that the relationship

$$E_{t-1} y^*_{t+1} = v^*_{t-1} \quad ,$$

holds, that is we have *perfect foresight.* Said differently, the stationary Nash solution satisfies the side condition of myopic foresight introduced in *section 2*, in spite of the fact that the two solutions (of *Theorem 1* and *Theorem 2*) are structurally different. [Compare (20) (or its stationary version) with (17)-(18).] This clearly implies that the Nash solution is disadvantageous to agent B (at least in the limit as $T \to \infty$), since it does not yield the best (optimum) solution obtainable under the side condition induced by the equilibrium solution itself. A reason for this *inefficient* behavior on the part of B is that in the analysis of *section 3* agent A is also an active player, whereas in *section 2* he was passive. Such features can be observed even in finite-horizon problems, as the following example demonstrates.

Numerical example 1. In our general formulation, let $T=0$, $E[y_0^2] =: \sigma_0$, and all other parameter values be unity. Then, the two solutions given in *Theorem 1* and *Theorem 2* and the corresponding values of expected costs and trajectory sequences can be computed to be as follows:

Theorem 1:

$$w_1 = \hat{\mu}_1(y_1) = -\frac{1}{2}y_1, \quad w_0 = \hat{\mu}_0(y_0) = -\frac{6}{7}y_0, \quad v_0 = \frac{1}{7}y_0, \tag{22a}$$

$$\hat{J}_A = \frac{5}{4}; \quad \hat{J}_B = \frac{5}{2} + \frac{6}{7}\sigma_0,$$

$$\hat{y}_1 = \frac{2}{7}y_0 + \epsilon_1; \quad \hat{y}_2 = \frac{1}{2}\hat{y}_1 + \epsilon_2.$$

Theorem 2:

$$w_1 = \mu^*_1(y_1) = -\frac{1}{2}y_1, \quad w_0 = \mu^*_0(y_0) = -\frac{3}{4}y_0, \quad v_0 = \gamma^*_0(y_0) = \frac{1}{4}y_0, \tag{22b}$$

$$J^*_A = \frac{5}{4}; \quad J^*_B = \frac{5}{2} + \frac{15}{16}\sigma_0,$$

$$y^*_1 = \frac{1}{2}y_0 + \epsilon_1; \quad y^*_2 = \frac{1}{2}y^*_1 + \epsilon_2.$$

A number of observations can be made in connection with this example:

1. In both cases above, we obtain *perfect foresight* (i.e. $v_0 = E_0 y_2$), but the corresponding trajectories are different. Even though (as we have seen earlier) the Nash solution does not generally enjoy perfect foresight for the finite-horizon case, here it does, mainly because the problem involves basically a single stage, thus eliminating the effect of spillover across consecutive periods.
2. Agent A incurs equal expected costs in both cases, whereas agent B does worse with the Nash solution. This is, of course, consistent with our earlier comments just preceding this example, which, even though were made in the context of the infinite-horizon problem, are equally valid here since the Nash solution satisfies the boundary condition (i.e. perfect foresight) of the myopic solution.

3. Since $\hat{\mu}_1 = \mu_1^*$ is a *universally optimal* policy for agent B at stage t=1, whichever equilibrium solution is adopted (even outside the two considered here) the trajectory will be given by

$$y_2 = \frac{1}{2}y_1 + \epsilon_2$$
$$y_1 = y_0 + v_0 + w_0.$$

Now, if we let $v_0 = E_0 y_2$, and attempt to solve for v_0 from the above equations, we first obtain (since $w_0 = \mu_0(y_0)$ is known to B for each fixed μ_0)

$$v_0 = E_0 y_2 = \frac{1}{2}E_0 y_1 = \frac{1}{2}y_0 + \frac{1}{2}v_0 + \frac{1}{2}w_0,$$

from which v_0 can be solved uniquely to give

$$v_0 = \gamma_0(y_0) = y_0 + w_0; \quad w_0 = \mu_0(y_0). \tag{$\circ$}$$

This shows that the actual choice for $v_0 = \gamma_0(y_0)$ (under perfect foresight) depends explicitly on B's policy μ_0, and the two solutions given above are two different manifestations of this dependence. Both (22*a*) and (22*b*) use ($\circ$) as a constraint, but while in (22*a*) J_B is minimized subject to ($\circ$), in (22*b*) the choices are determined by the Nash solution of a game played between the two agents at time t=0. One could envision other scenarios between the two agents which would lead to still different choices for μ_0 (and thereby γ_0), but in all cases the resulting expected cost to A will be the constant 5/4, independent of μ_0 and σ_0. ◇

We now conclude this section with a second example, which is an extended version of the previous example with an additional stage. It will serve to demonstrate some additional features of the solution given in *Theorem 2*.

Numerical example 2. In the general formulation, let T=1, and all parameter values be unity. Then, the unique Nash equilibrium solution (as presented in *Theorem 2*) can be computed to be as follows:

$$\begin{aligned}
&w_2 = \mu_2^*(y_2) = -\frac{1}{2}y_2, \quad w_1 = \mu_1^*(y^1) = -\frac{3}{8}y_1 - 0.190476y_0,\\
&w_0 = \mu_0^*(y_0) = -0.746032y_0;\\
&v_1 = \gamma_1^*(y^1) = -\frac{3}{8}y_1 + 0.31746y_0, \quad v_0 = \gamma_0^*(y_0) = 0.253968y_0.
\end{aligned} \tag{23}$$

The corresponding equilibrium trajectory is generated by

$$\begin{aligned}
y_1^* &= 0.5079366y_0 + \epsilon_1\\
y_2^* &= 0.25y_1^* + 0.126984y_0 + \epsilon_2\\
y_3^* &= 0.5y_2^* + \epsilon_3 \quad ,
\end{aligned}$$

from which it follows that $E_1 y_3^* = 0.125y_1^* + 0.063492y_0 \not\equiv \gamma_1^*(y_1^*, y_0)$; that is, the solution does not lead to perfect foresight at time t=1. However, $E_0 y_2^* = 0.253968y_0 \equiv \gamma_0^*(y_0)$; that is, there is perfect foresight at t=0. This latter result is not a feature of this example only, but holds for the general solution of *Theorem 2* (even though it may be rather difficult to prove algebraically). Through an indirect reasoning that follows the proof of *Theorem 2*, as given in the *Appendix*, one can conclude that $E_0 y_2^* = \gamma_0^*(y_0)$ is a genuine property of the general Nash solution since, at the initial stage, the variable v_0 minimizes an expression that is a perfect square in y_2 (see $(A.1)$) and there is no spillover effect. ◇

4. Some Extensions

A first extension of the results presented in *section 2* and *section 3* would be to the more general case where the reference trajectory is not *zero* and the cost function (5b) contains additional (time-varying) weights on the deviation from the desired trajectory (*i.e.* the first term). The reason why we have not included this in our presentation here is because such an extension does not entail anything conceptually new, while requiring some additional notation which would have complicated the resulting expressions considerably. The gist of the results for the nonzero reference trajectory case is that the statements of both *Theorem 1* and *Theorem 2* remain essentially intact, with the only difference being that now each policy includes an additive (bias) term which depends linearly on the desired reference trajectory. The existence conditions in both cases are identical to the earlier ones. For the case when there is a time-varying weight in the first term of (5b), the results again remain intact, with only the additive term 1 in (8a) replaced by this new weight and Q in (13a) adjusted accordingly.

A second extension would be to the class of problems where the agents do not have direct access to the trajectory $\{y_t\}$, but rather acquire common noisy measurements $\{z_t\}$, as defined by (1b), where now $\eta_t = z^t$. Towards studying this extension, let us assume that $\{\epsilon_t\}$ and $\{\xi_t\}$ are sequences of independent Gaussian zero-mean random variables with variances $var(\epsilon_t) =: \varphi > 0$, $var(\xi_t) =: \varsigma_t > 0$, and that they are independent of y_0 which is also a Gaussian zero-mean random variable, with variance σ_0. Then, in the formulation of *section 2*, we interpret the operator E_t as the conditional expectation $E\{\cdot|z^t, w^{t-1}\}$. Note that here we have replaced $\eta_t = z^t$ with $\tilde{\eta}_t := (z^t, w^{t-1})$, without any loss of generality, since w^{t-1} is measurable with respect to z^{t-1}. Now, letting $\hat{y}_t := E_t y_t$, it is a standard result (see, for example, *Bertsekas (1987)* or *Kumar and Varaiya (1986)*) that $\hat{y}_t$ is generated by the Kalman filter equations:

$$\begin{aligned} \hat{y}_{T+2} &= a\hat{y}_{T+1} + cw_{T+1} + [\hat{\sigma}_{T+2}/(\hat{\sigma}_{T+2} + \varsigma_{T+2})]r_{T+2}, \\ \hat{y}_{t+1} &= a\hat{y}_t + bE_t y_{t+2} + cw_t + [\hat{\sigma}_{t+1}/(\hat{\sigma}_{t+1} + \varsigma_{t+1})]r_{t+1}, \quad t \le T; \quad \hat{y}_{-1} = 0, \end{aligned} \tag{24a}$$

$$r_{t+1} := z_{t+1} - a\hat{y}_t - bE_t y_{t+2} - cw_t, \tag{24b}$$

$$\hat{\sigma}_{t+1} = [a^2\varsigma_t/(\hat{\sigma}_t + \varsigma_t)]\hat{\sigma}_t + \varphi_{t+1}, \quad \hat{\sigma}_0 = \sigma_0, \tag{24c}$$

where $\{r_t\}$ is a sequence of independent Gaussian random variables, known as the *innovation* sequence. In writing down these relationships, we have made explicit use of the fact that both $E_t y_{t+2}$ and w_t are z^t-measurable.

Now note that the error sequence $\{e_t\}$, $e_t := y_t - \hat{y}_t$, is generated by

$$e_{t+1} = ae_t + \epsilon_{t+1} - [\hat{\sigma}_{t+1}/(\hat{\sigma}_{t+1} + \varsigma_{t+1})]r_{t+1}; \quad e_0 = 0, \tag{Δ}$$

and that $E_t e_{t+n} = 0$ for all $n \ge 1$. In view of this last property,

$$E_t y_{t+2} = E_t\hat{y}_{t+2} + E_t e_{t+2} = E_t\hat{y}_{t+2},$$

and hence (24a) can be rewritten as

$$\begin{aligned} \hat{y}_{T+2} &= a\hat{y}_{T+1} + cw_{T+1} + [\hat{\sigma}_{T+2}/(\hat{\sigma}_{T+2} + \varsigma_{T+2})]r_{T+2}, \\ \hat{y}_{t+1} &= a\hat{y}_t + bE_t\hat{y}_{t+2} + cw_t + [\hat{\sigma}_{t+1}/(\hat{\sigma}_{t+1} + \varsigma_{t+1})]r_{t+1}, \quad t \le T; \quad \hat{y}_{-1} = 0. \end{aligned} \tag{25}$$

Furthermore, since $y_t = \hat{y}_t + e_t$, and $\hat{y}_t$ is orthogonal to e_t, the counterpart of (7*a*) for the noisy case would be:

$$F_s^{T+1} = \sum_{t=s}^{T+1} E\{\hat{y}_{t+1}^2 + kw_t^2\}\rho_B^{t-s} + \sum_{t=s}^{T+1} E\{e_{t+1}^2\}\rho_B^{t-s}, \tag{26}$$

where the second summation term does not enter the optimization, since the sequence $\{e_t\}$ generated by $(\triangle)$ is independent of the control sequence $\{w_t\}$. Hence, the problem faced by agent B is the minimization of the first term of (26) subject to the dynamics (25), where $w_t = \mu_t(\hat{y}^t)$, which is compatible with the original information $\tilde{\eta}_t = (y^t, w^{t-1})$ since $\hat{y}_t$ is generated by $(\tilde{\eta}_{t-1}, w_{t-1})$. Then, the problem is identical with the perfect information case (apart from a change of notation), in view of the fact that $\{r_t\}$ is a zero-mean independent sequence, playing the role of $\{\epsilon_t\}$ in (7*b*). This shows that the problem (with myopic forecast) features *certainty equivalence*, making the statement of *Theorem 1* valid also in the noisy case, with only y_t replaced by $\hat{y}_t$, and (9*c*) including an additional positive term due to the second term of (26). The following theorem summarizes this result.

Theorem 3. *Let Condition 1 be satisfied. Then, the dynamic policy optimization problem with myopic forecast, as formulated in section 2 but with common noisy measurements* (1*b*) *for both agents, admits the unique solution*

$$w_t = \hat{\mu}_t(y_t) = g_t\hat{y}_t, \quad 0 \le t \le T+1, \tag{27a}$$

with the corresponding forecast policy given by

$$v_t = E_t y_{t+2} = \frac{(a + cg_{t+1})(a + cg_t)}{\nu_{t+1}\nu_t}\hat{y}_t, \tag{27b}$$

where $\{\hat{y}_t\}$ *is generated by* (25), *and* $\{g_t\}$, $\{\nu_t\}$ *are as defined by* (8*c*) *and* (8*b*), *respectively.*
◇

Hence, for the noisy case, certainty equivalence holds under myopic forecast, and the statement of *Theorem 1* basically remains intact. For *Theorem 2*, however, there is no direct counterpart, and derivation of the Nash equilibrium solution is quite a nontrivial task. We will not pursue this extension here, since presenting the full details of the derivation of the Nash equilibrium solution would at least double the length of the present paper. What we can say at this point, however, is that (guided by the results presented in *Başar (1978b)* for a linear-quadratic nonzero-sum dynamic game with a different type of an information pattern and a different type of a cost function for one of the agents) the problem will generically admit a unique Nash equilibrium solution, linear in the available common information. This solution will not satisfy the certainty equivalence or separation principle of stochastic control, and thus will have no relationship with the solution presented in *Theorem 2*. The following numerical example (which is the "noisy" version of the second example of *section 3*) should serve to corroborate this claim and to give some indication as to the intricacies involved in the derivation of the general solution.

Numerical example 3. Consider the second numerical example of *section 3*, but with noisy measurement (1*b*) for both agents, and with all parameter values (including the noise variances) equal to unity. Hence, the cost functions are

$$\begin{aligned} J_A &= E\{(v_1 - y_3)^2 + (v_0 - y_2)^2\} \\ J_B &= E\{y_3^2 + w_2^2 + y_2^2 + w_1^2 + y_1^2 + w_0^2\}, \end{aligned} \tag{28}$$

and the dynamic constraints are

$$\begin{aligned} y_3 &= y_2 + w_2 + \epsilon_3 \\ y_2 &= y_1 + v_1 + w_1 + \epsilon_2 \\ y_1 &= y_0 + v_0 + w_0 + \epsilon_0, \end{aligned} \tag{29}$$

where $w_2 = \mu_2(z^2)$, $w_1 = \mu_1(z^1)$, $w_0 = \mu_0(z_0)$, $v_1 = \gamma_1(z^1)$, $v_0 = \gamma_0(z_0)$; $z_0 = y_0 + \xi_0$, $z_1 = y_1 + \xi_1$ and $z_2 = y_2 + \xi_2$. The first significant difference between the perfect and the noisy measurement cases appears in the construction of the best μ_2, which now depends explicitly on (μ_1, μ_0) and (γ_1, γ_0). [Recall that in the perfect measurement case covered by *Theorem 2*, there was a universally optimal policy for agent B at the terminal stage of the game.] With the quadruple $(\gamma_1, \gamma_0; \mu_1, \mu_0)$ fixed, say at (γ, μ), the minimization of J_B with respect to μ_2 becomes a standard quadratic optimization problem,

$$\min_{w_2} E\{(y_2 + w_2 + \epsilon_3)^2 + w_2^2 | z^2, \gamma, \mu\},$$

whose unique solution is

$$w_2 = \mu_2(z^2, \gamma, \mu) = -\frac{1}{2} E[y_2 | z^2, \gamma, \mu] =: \frac{1}{2} \hat{y}_{2\gamma\mu}. \tag{30}$$

Here $\hat{y}_{2\gamma\mu}$ is generated by the Kalman filter:

$$\begin{aligned} \hat{y}_{2\gamma\mu} &= \hat{y}_{1\gamma_0\mu_0} + \gamma_1(z^1) + \mu_1(z^1) + \frac{8}{13}(z_2 - \hat{y}_{1\gamma_0\mu_0} - \gamma_1(z^1) - \mu_1(z^1)) \\ \hat{y}_{1\gamma_0\mu_0} &= \hat{y}_0 + \gamma_0(z_0) + \mu_0(z_0) + \frac{3}{5}(z_1 - \hat{y}_0 - \gamma_0(z_0) - \mu_0(z_0)) \\ \hat{y}_0 &= \frac{1}{2} z_0, \end{aligned} \tag{31}$$

which depends on (γ, μ) partly directly and partly through $\hat{y}_{1\gamma_0\mu_0} := E[y_1 | z^1, \gamma_0, \mu_0]$. To obtain the pair (γ, μ) that is in Nash equilibrium with (30), we follow a procedure quite analogous (in principle) to the one followed in the proof of uniqueness (for *Theorem 2*) in the *Appendix*, geared towards obtaining a (unique) stagewise equilibrium. Accordingly, the derivation involves the solution of two static games, one at t=1 and the other one at t=0. To characterize the static game at t=1, we substitute (31) into (28), eliminate the intermediate variables and take expectation over the statistics of ϵ_3, ϵ_2 and ξ_2, to arrive at the reduced conditional (on z^1) cost functions:

$$\begin{aligned} J_A^1 = E\{\frac{1}{169}[4v_1 - 9(y_1 + w_1) + \frac{5}{2}\hat{y}_1 + \frac{5}{2}(\gamma_1 + \mu_1)]^2 \\ + [v_0 - y_1 - v_1 - w_1]^2 | z^1\}, \end{aligned}$$

$$\begin{aligned} J_B^1 = E\{\frac{1}{169}[9(y_1 + v_1 + w_1) - \frac{5}{2}\hat{y}_1 - \frac{5}{2}(\gamma_1 + \mu_1)]^2 \\ + \frac{1}{676}[5(\hat{y}_1 + \gamma_1 + \mu_1) + 8(y_1 + v_1 + w_1)]^2 + (y_1 + v_1 + w_1)^2 + w_1^2 | z^1\}. \end{aligned}$$

In the above, we have made notational simplifications by suppressing the (γ_0, μ_0)-dependence of $\hat{y}_1$ and the arguments of (γ_1, μ_1). This is clearly a static game in the pair (v_1, w_1), and its Nash solution can be obtained for each fixed (γ_1, μ_1) and (γ_0, μ_0), where we take $v_0 = \gamma_0(z_0)$.

Differentiating J_A^1 with respect to v_1 and J_B^1 with respect to w_1, and setting the resulting expressions equal to zero after conditioning on z^1, we find that the Nash condition is satisfied and there exists a unique solution to the pair of equations, linear in γ_1, μ_1, $\hat{y}_1$ and γ_0. Now requiring consistency in the solution (as in the proof of uniqueness for *Theorem 2* in the *Appendix*), we set $v_1 = \gamma_1(z^1)$, $w_1 = \mu_1(z^1)$, and solve the resulting pair of linear equations (in v_1 and w_1) uniquely, to arrive at the policies:

$$v_1 = \gamma_1(z^1, \gamma_0, \mu_0) = -0.523810\hat{y}_1 + 1.547619\gamma_0(z_0), \tag{32a}$$

$$w_1 = \mu_1(z^1, \gamma_0, \mu_0) = -0.285714\hat{y}_1 - 0.928571\gamma_0(z_0). \tag{32b}$$

Note that here γ_0 is yet to be determined.

To complete the solution, we next formulate the game at t=0, by substituting (30) and (32) into (28), again eliminating the intermediate variables and averaging over the statistics of the random variables involved, to obtain the reduced conditional (on z_0) cost functions:

$$\begin{aligned} J_A^0 = E\{&(0.619048\hat{y}_0 + 1.208790\gamma_0 - 0.029304\mu_0 + 0.648352(v_0 + w_0))^2 \\ &+ (0.485714v_0 + 0.295238\hat{y}_0 - 0.514286w_0 + 0.190477\gamma_0 + 0.809524\mu_0)^2 | z_0\}, \end{aligned}$$

$$\begin{aligned} J_B^0 = E\{&\frac{1}{4}(0.619048\gamma_0 - 1.194139\hat{y}_1 + 1.384615y_1)^2 \\ &+ \frac{1}{4}(0.619048\gamma_0 - 0.424909\hat{y}_1 + 0.615385y_1)^2 + (0.619048\gamma_0 - 0.809524\hat{y}_1 + y_1)^2 \\ &+ (0.285714\hat{y}_1 + 0.928571\gamma_0)^2 + y_1^2 + w_0^2 | z_0\}, \end{aligned}$$

where both y_1 and $\hat{y}_1$ depend on (v_0, w_0), the latter through z_1, as given in (31).

The procedure here is the same as at t=1: First obtain the Nash solution of (J_A^0, J_B^0) in terms of (γ_0, μ_0), then require consistency (v_0=$\gamma_0(z_0)$,w_0=$\mu_0(z_0)$) and solve for (v_0, w_0) from the resulting equations, which will lead to policies whose argument is z_0. At each step the uniqueness condition is met, and thus the procedure yields the unique Nash equilibrium policies (at t=0):

$$v_0 = \gamma_0^*(z_0) = +0.248227\hat{y}_0, \tag{33a}$$

$$w_0 = \mu_0^*(z_0) = -0.751773\hat{y}_0. \tag{33b}$$

These policies are finally used in (32) and (31) to complete the characterization of the Nash equilibrium policies:

$$v_1 = \gamma_1^*(z^1) = -0.523810\hat{y}_1^* + 0.384161\hat{y}_0, \tag{34a}$$

$$w_1 = \mu_1^*(z^1) = -0.285714\hat{y}_1^* - 0.230496\hat{y}_0, \tag{34b}$$

$$w_2 = \mu_2^*(z^2) = -0.5\hat{y}_2, \tag{34c}$$

where

$$\begin{aligned} \hat{y}_2^* &= 0.059102\hat{y}_0 + 0.073260\hat{y}_1^* + 0.615385z_2 \\ \hat{y}_1^* &= 0.198582\hat{y}_0 + 0.6z_1 \\ \hat{y}_0 &= 0.5z_0. \end{aligned} \tag{35}$$

An equivalent representation for γ_1^* in (34*a*) would be

$$v_1 = \gamma_1^*(z^1) = -0.523810\hat{y}_1^* + 1.547619v_0,$$

which shows explicit dependence on v_0. Note that the policies (34) are different from their counterparts in the noise-free case (*i.e.* (23)), thus corroborating our earlier remark that the "noisy version" does not feature certainty equivalence.

The equilibrium trajectory corresponding to the unique Nash solution is generated by

$$\begin{aligned} y_3^* &= y_2^* - 0.5\hat{y}_2^* + \epsilon_3 \\ y_2^* &= y_1^* - 0.809524\hat{y}_1^* + 0.153664\hat{y}_0 + \epsilon_2 \\ y_1^* &= y_0 - 0.503546\hat{y}_0 + \epsilon_1. \end{aligned}$$

Using these, it is easy to check that, as in the second example of *section 3*, $E_1 y_3^* \not\equiv \gamma_1^*(z^1)$, while $E_0 y_2^* \equiv \gamma_0^*(z_0)$, which shows that the Nash solution could lead to perfect foresight at the initial stage, even in the noisy case. As we will discuss in a companion paper, this turns out to be a general property of the Nash solution for the "noisy version" of the problem of *section 3.* ◇

Appendix

In this appendix, we first complete the proof of the existence part of *Theorem 2* by showing that the policy $(\star\star)$ given there indeed solves agent A's optimization problem. Subsequently, we establish the uniqueness of the Nash solution presented in *Theorem 2.*

Existence. The optimization problem faced by agent A is the minimization of J_0^T, where

$$J_s^T = \sum_{t=s}^{T} E\{(v_t - y_{t+2})^2\}\rho_A^{t-s},$$

under the constraints

$$\begin{gathered} y_{T+2} = [ak/(k+c^2)]y_{T+1} + \epsilon_{T+2} \\ y_{t+1} = (a + c\tilde{\alpha}_t)y_t + c\tilde{\beta}_t\tilde{v}_{t-1} + bv_t + \epsilon_{t+1}, \quad 1 \le t \le T \\ y_1 = (a + c\tilde{\alpha}_0)y_0 + \epsilon_1; \\ \tilde{v}_t = \alpha_t y_t + \beta_t \tilde{v}_{t-1}, \quad v_t = \gamma_t(y^t). \end{gathered}$$

We now claim that, for a general t,

$$\min_{\{\gamma_s\}_{s=t}^{T}} J_t^T = \min_{\gamma_{t+1},\gamma_t} E\{\rho_A n_{t+1}(v_{t+1} - m_{t+1}y_{t+2} - \tilde{m}_{t+1}\tilde{v}_{t+1})^2 + (v_t - y_{t+2})^2\} + q_t, \quad (A.1)$$

where $\{q_t\}$ is a sequence depending only on the variances of the additive stochastic terms ϵ_t, $t \le T + 2$. Under the validity of this assertion, the optimal policy at time t is obtained by minimizing the following quantity with respect to the scalar variable $v_{t+1} =: v$, for each fixed y^{t+1}:

$$\begin{aligned} E\{\rho_A n_{t+1}[v - m_{t+1}(a + c\tilde{\alpha}_{t+1})y_{t+1} - m_{t+1}c\tilde{\beta}_{t+1}\tilde{v}_t - m_{t+1}bv - \tilde{m}_{t+1}\alpha_{t+1}y_{t+1} \\ - \tilde{m}_{t+1}\beta_{t+1}\tilde{v}_t]^2 + [(a + c\tilde{\alpha}_{t+1})y_{t+1} + c\tilde{\beta}_{t+1}\tilde{v}_t + bv - v_t]^2 | y^{t+1}\}. \end{aligned} \quad (A.2)$$

Being quadratic and strictly convex (in v), this optimization problem admits a unique solution (for each fixed $y_{t+1}, \tilde{v}_t, v_t$), given by

$$v_{t+1} = \gamma_{t+1}(y^{t+1}) = \hat{\alpha}_{t+1}y_t + \hat{\beta}_{t+1}v_t + \bar{\beta}_{t+1}\tilde{v}_t, \quad (A.3a)$$

for $0 \leq t \leq T-1$, and at the initial stage by

$$v_0 = \gamma_0(y_0) = \hat{\alpha}_0 y_0, \tag{A.3b}$$

where

$$\widehat{\alpha}_t = \frac{1}{b^2 + (1-bm_t)^2 \rho_A n_t}[\rho_A n_t(1-bm_t)(\widetilde{m}_t \alpha_t + (a + c\widetilde{\alpha}_t)m_t) - b(a + c\widetilde{\alpha}_t)]$$

$$\widehat{\beta}_t = \frac{b}{b^2 + (1-bm_t)^2 \rho_A n_t}$$

$$\bar{\beta}_t = \frac{1}{b^2 + (1-bm_t)^2 \rho_A n_t}[\rho_A n_t(1-bm_t)(\widetilde{m}_t \beta + c\widetilde{\beta}_t m_t) - bc\widetilde{\beta}_t].$$

As we have discussed earlier (in the proof of the existence part of *Theorem 2*), substitution for α_t and β_t (from (11*a*) and (12*a*), respectively) into the three expressions above, leads to the equivalences $\hat{\alpha}_t \equiv \alpha_t$ and $\hat{\beta}_t + \bar{\beta}_t \equiv \beta_t$. Hence, the optimal solution (*A*.3) admits the equivalent representation

$$\begin{aligned} v_t &= \alpha_t y_t + \hat{\beta}_t v_{t-1} + (\beta_t - \hat{\beta}_t)\tilde{v}_{t-1}, \quad && 1 \leq t \leq T \\ &= \alpha_0 y_0 \qquad\qquad , && t = 0. \end{aligned} \tag{A.4}$$

We now turn to verification of the structural form (*A*.1). The result trivially holds for t=T, with $m_T = ak/(k+c^2), \widetilde{m}_T = 0$. Let us therefore assume the validity of the assertion for t+1 and prove it for t. Towards this end, we substitute (*A*.4), with t replaced by t+1, into (*A*.2), and arrive (after some rather tedious algebra) at an expression which is a perfect square in v_t, y_{t+1} and $\tilde{v}_t$:

$$E\{n_t(v_t - m_t y_{t+1} - \widetilde{m}_t \tilde{v}_t)^2 | y^{t+1}\}. \tag{A.5}$$

Here m_t, $\widetilde{m}_t$ and n_t are defined in terms of m_{t+1}, $\widetilde{m}_{t+1}$ and n_{t+1} as in (10*a*) through (10*c*). [In fact, it is not difficult to see that the resulting cost should be a perfect square, because (*A*.2) can be made equal to zero by appropriately choosing v_t and v_{t+1}. With this observation, it then remains to find the three coefficients n_t, m_t and $\widetilde{m}_t$.] Now, since the minimum of (*A*.2) over v_{t+1} is equal to (*A*.5), we have

$$\begin{aligned} \min_{\{\gamma_s\}_{s=t-1}^T} J_{t-1}^T &= \min_{\gamma_t, \gamma_{t-1}} E\{(v_{t-1} - y_{t+1})^2 + \rho_A \min_{\{\gamma_s\}_{s=t+1}^T} J_t^T\} \\ &= \min_{\gamma_t, \gamma_{t-1}} E\{(v_{t-1} - y_{t+1})^2 \rho_A n_t (v_t - m_t y_{t+1} - \widetilde{m}_t \widetilde{v}_t)^2\} \\ &\qquad + \rho_A[q_t + (1 + \rho_A n_{t+1} m_{t+1}^2) var(\epsilon_{t+2})], \end{aligned}$$

which is in the same form as (*A*.1), with

$$q_{t-1} := \rho_A[q_t + (1 + \rho_A n_{t+1} m_{t+1}^2) var(\epsilon_{t+2})].$$

This then completes the proof of optimality of (⋆⋆) in the proof of the existence part of *Theorem 2*.

Uniqueness. It is a well-known fact that dynamic games could admit nonunique Nash equilibria, with each such equilibrium leading to a different cost pair which are in general incomparable (see, for example, *Başar and Olsder (1982)*). Thus, "uniqueness" is an important question to pose, if the proposed equilibrium is to be of value. As we have discussed extensively in earlier papers (for example, *Başar(1976)*, *Başar(1977)*), the main source of nonuniqueness in Nash equilibria is the so-called *informational nonuniqueness* which arises if each agent, in his one-sided optimization, has the freedom of choosing different *representations* of the same policy. What we prove in the sequel is that for the game problem covered by *Theorem 2* there is no informational nonuniqueness, and the structural form (17)-(18) is the only form in which a Nash equilibrium can exist. Furthermore, we show that structural uniqueness is guaranteed under *Condition 2*. In the proof, we will not explicitly derive the expressions for this unique Nash solution, since we have already shown in the first part of the proof that (17)-(18) exists as a Nash equilibrium.

Towards devising a proof for uniqueness, we first introduce two *generic* functions $quad(\cdot)$ and $lin(\cdot)$, where

$$quad(\cdot) = \text{a quadratic function of its arguments}$$
$$lin(\cdot) = \text{a linear function of its arguments.}$$

Furthermore, we introduce a class of *nested subgames* $\{\mathrm{G}_s\}$, parameterized by s, each one being a replica of the original game but defined on a shorter time interval, $[s, T+1]$, $0 \le s \le T+1$. More precisely, for the subgame G_s, the cost functions are defined by (5*a*)-(5*b*) with the lower limits changed to $t{=}s{-}1$, and with the action variables being $v_s^T := (v_T, .., v_{s+1}, v_s)$ for A, and $w_s^{T+1} := (w_{T+1}, .., w_{s+1}, w_s)$ for B, where $v_t = \mu_t(y^t)$, $w_t = \gamma_t(y^t)$, and a similar convention as above applying to the policy variables μ_s^{T+1}, γ_s^T. To be consistent with this convention, for $s{=}0$ we extend the limit of the summation to $t = -1$ in both J_A and J_B, by adding *zero* as the incremental cost term at $t = -1$. Now let $(\tilde{\gamma} := \tilde{\gamma}_0^T, \tilde{\mu} := \tilde{\mu}_0^{T+1})$ be a Nash equilibrium solution for the original game (G_0), such as the one given in *Theorem 2*. Then, it is a well-known property of the Nash solution (called the *stagewise equilibrium* property) that for any s, the truncated version of these policies, $(\tilde{\gamma}_s^T, \tilde{\mu}_s^{T+1})$, constitutes a Nash equilibrium solution for G_s, with the past policies $(\gamma_0^{s-1}, \mu_0^{s-1})$ fixed at $(\tilde{\gamma}_0^{s-1}, \tilde{\mu}_0^{s-1})$.

We now develop a procedure for studying the uniqueness of the solutions of these individual subgames. First consider the case $s = T+1$, where G_{T+1} is not really a game but a one-sided optimization problem for agent B, since only B is active at $t = T+1$. Then, clearly the solution is unique, and is given by the second line in (*18a*). Note that this solution is both informationally and structurally unique (regardless of the past policy choices), the former being due to our assumption in *section 1* on the structure of the probability distribution of the additive system noise. Hence, in the study of the second game in the sequence, G_T, we can take μ_{T+1} as in (18*a*), without any loss of generality. Accordingly, substituting this μ_{T+1}, say μ_{T+1}^*, into both J_A and J_B, eliminating the intermediate variables using the evolution equation (4) and averaging over the statistics of the random variables by employing their independence property, we arrive at the structural forms

$$\begin{aligned} cost_A(\mathrm{G}_T) &= quad(y_T, v_T, w_T, v_{T-1}) \\ cost_B(\mathrm{G}_T) &= quad(y_T, v_T, w_T) + quad(w_{T-1}^2), \end{aligned} \qquad (A.6)$$

which are the costs incurred to A and B, respectively, conditioned on the information available at time T, *i.e.* $\eta_T{=}y^T$. Since the first cost shows explicit dependence on v_{T-1}, we fix $v_{T-1} = \gamma_{T-1}(y^{T-1})$, and solve for the Nash equilibrium of the resulting static game. Because of the quadratic structure of the cost functions, the Nash solution, if it exists, will be linear in the pair (y_T, v_{T-1}); furthermore it will be (structurally) unique under conditions not depending on y_T and v_{T-1}, and *Condition 2* precisely serves this purpose. Hence, the static game defined by (*A.6*) admits a unique Nash solution, for each fixed γ_{T-1}, given by

$$v_T = \tilde{\gamma}_T(y^T) \equiv lin(y_T, v_{T-1}) \qquad (A.7a)$$

$$w_T = \tilde{\mu}_T(y^T) \equiv lin(y_T, v_{T-1}), \tag{A.7b}$$

where $v_{T-1} = \gamma_{T-1}(y^{T-1})$. The linear functions here are precisely the ones given in (17) and (18), with v_{T-1} in the latter case replaced by $\tilde{v}_{T-1}$. The solution is also unique *representationwise* since, because of our nonsingular statistics assumption on the probability distributions of the random variables involved, y_T cannot be expressed in terms of the past values of the trajectory *almost surely* (which would have been possible in a purely deterministic problem). We now note that the complete (unique) solution to subgame G_T is (*A*.7) along with μ^*_{T+1} which was the unique solution (for agent B) to subgame G_{T+1}.

The next game in the sequence, G_{T-1}, involves the the action variables (v_T, v_{T-1}) for agent A and (w_{T+1}, w_T, w_{T-1}) for agent B. Since every Nash equilibrium is necessarily a stagewise equilibrium and since the unique (linear) Nash solution of G_T does not depend structurally on v_{T-1} and w_{T-1}, it follows that every Nash equilibrium for G_{T-1} should match with that of G_T for policies μ_{T+1}, μ_T and γ_T. Hence, the equilibrium solution of G_{T-1} will be nonunique only if the last components of the policy sequences, $(\gamma_{T-1}, \mu_{T-1})$, are nonunique at equilibrium. Towards a study of this, we substitute the solution of G_T into J_A and J_B, with v_{T-1} in (*A*.7*b*) replaced by a general function of y^{T-1}, say $\psi_{T-1}(y^{T-1})$, since B does not have direct access to v_{T-1}. [It is important to note at this point that if B had direct access to v_{T-1}, the solution would have been informationally nonunique, for reasons discussed extensively in *Başar (1978a)* for a different class of such games.] Now, after eliminating the intermediate variables and averaging over the stochastic variables, we arrive at the following *reduced* costs for G_{T-1}, conditioned on the common information available at time $T-1$, *i.e.* y^{T-1}:

$$cost_A(\mathrm{G}_{T-1}) = quad(y_{T-1}, v_{T-1}, w_{T-1}, v_{T-2}, \psi_{T-1}(y^{T-1}))$$
$$cost_B(\mathrm{G}_{T-1}) = quad(y_{T-1}, v_{T-1}, w_{T-1}, \psi_{T-1}(y^{T-1})).$$

Here, in addition to the unknown (but fixed) function ψ_{T-1}, we also have $v_{T-2} = \gamma_{T-2}(y^{T-2})$ fixed by an arbitrary choice of γ_{T-2}. Under an appropriate condition which is independent of ψ_{T-1} and γ_{T-2} (which is also guaranteed by *Condition 2*), this static game admits a unique equilibrium for each fixed ψ_{T-1} and γ_{T-2}:

$$v_{T-1} = \hat{\gamma}_{T-1}(y_{T-1}, v_{T-2}, \psi_{T-1}(y_{T-1})) \equiv lin(y_{T-1}, v_{T-2}, \psi_{T-1}(y_{T-1})) \tag{A.8a}$$

$$w_{T-1} = \hat{\mu}_{T-1}(y_{T-1}, v_{T-2}, \psi_{T-1}(y_{T-1})) \equiv lin(y_{T-1}, v_{T-2}, \psi_{T-1}(y_{T-1})), \tag{A.8b}$$

where $v_{T-2}=\gamma_{T-2}(y^{T-2})$. Next, we impose consistency in the solution for each fixed γ_{T-2}, which requires that $\hat{\gamma}_{T-1} \equiv \psi_{T-1}$. Using this side condition in (*A*.8*a*), we arrive at

$$v_{T-1} = lin(y_{T-1}, v_{T-2}, v_{T-1})$$

which, being linear admits the unique solution (for each fixed y^{T-1} and γ_{T-2})

$$v_{T-1} = \tilde{\gamma}_{T-1}(y_{T-1}, v_{T-2}) \equiv lin(y_{T-1}, v_{T-2}), \tag{A.9a}$$

under a nonsingularity condition which is met under *Condition 2*. Letting $\psi_{T-1} = \tilde{\gamma}_{T-1}$ in (*A*.8*b*), we finally obtain for w_{T-1} (for each fixed γ_{T-2}):

$$w_{T-1} = \tilde{\mu}_{T-1}(y_{T-1}, v_{T-2}) \equiv lin(y_{T-1}, v_{T-2}) \tag{A.9b}$$

This then completes the verification of the uniqueness of the solution of G_{T-1}, for each fixed γ_{T-2}. Note that the complete solution to G_{T-1} is given by μ^*_{T+1}, (*A*.7) and (*A*.9), with

v_{T-1} in (*A.7b*) replaced by the expression in (*A.9a*). Here we could also have expressed (*A.7a*) in terms of y^T, instead of (y_T, v_{T-1}), by substituting for v_{T-1} from (*A.9a*), but this is not necessary since agent A does have access to his past decision value, and enriching his information set by also including past decision values does not lead to informational nonuniqueness.

The important observation here is that, for each fixed γ_{T-2}, the solution of subgame G_{T-1} (to be denoted $(\tilde{\gamma}_T, \tilde{\gamma}_{T-1}; \tilde{\mu}_{T+1}, \tilde{\mu}_T, \tilde{\mu}_{T-1})$) is structurally unique, with each strategy being linear in its arguments. More precisely, we have $\tilde{\gamma}_T$ linear in (y_T, v_{T-1}), $\tilde{\gamma}_{T-1}$ linear in (y_{T-1}, v_{T-2}), $\tilde{\mu}_{T+1}$ linear in y_{T+1}, $\tilde{\mu}_T$ linear in $(y_T, y_{T-1}, \gamma_{T-2}(y^{T-2}))$ and $\tilde{\mu}_{T-1}$ linear in $(y_{T-1}, \gamma_{T-2}(y^{T-2}))$. Furthermore, the solution is informationally unique because of the nonsingular statistics of the additive noise in the dynamics (4). Then, in the construction of the Nash solution for subgame G_{T-2}, we first substitute for $(\gamma_T, \gamma_{T-1}; \mu_{T+1}, \mu_T, \mu_{T-1})$ from the unique solution of G_{T-1}, with γ_{T-2} replaced by a general function ψ_{T-2}, as in the construction of the solution for G_{T-1}. Repeating the same procedure as in G_{T-1}, we can obtain a linear stagewise Nash solution for G_{T-2} for each fixed γ_{T-3}, whose uniqueness is again guaranteed by *Condition 2*. Following this procedure in retrograde time, we find that for each s, the subgame G_s admits a unique stagewise equilibrium (for each fixed γ_{s-1}), linear in the available information as well as in γ_{s-1}. Since γ_{-1} is trivially zero, the process halts at s=0, leading to the conclusion that the game G_0 admits a unique stagewise equilibrium, linear in the common information available to the agents. This then establishes uniqueness of the Nash solution of the original problem (which is identical with G_0), since every Nash equilibrium is a stagewise equilibrium and we have already proven that the game admits at least one Nash equilibrium.

We conclude this *Appendix* by pointing to the fact that the above procedure would have been an alternative method for the construction of the Nash solution given in *Theorem 2*, but alone it would not be sufficient, since a stagewise equilibrium need not be a Nash equilibrium. ◇

References

Barro, R. J. (1976) , "Rational Expectations and the Role of Monetary Policy", *Journal of Monetary Economics*, vol.2, pp.1-33.

Başar, T. (1976) , "On the Uniqueness of the Nash Solution in Linear-Quadratic Differential Games", *International Journal of Game Theory*, vol.5, no.2/3, pp.65-90.

Başar, T. (1977) , "Informationally Nonunique Equilibrium Solutions in Differential Games", *SIAM Journal on Control*, vol.15, no.4, pp.636-660.

Başar, T. (1978a) , "Decentralized Multicriteria Optimization of Linear Stochastic Systems", *IEEE Transactions on Automatic Control*, vol.AC-23, no.2, pp.233-243.

Başar, T. (1978b) , "Two-Criteria LQG Decision Problems with One-Step Delay Observation Sharing Pattern", *Information and Control*, vol.32, no.1, pp.21-50.

Başar, T. (1987) , "Some Thoughts on Rational Expectations Models, and Alternate Formulations", invited contribution to a special issue of *Computer and Mathematics with Applications*, to appear in 1988/1989.

Başar, T. and G. J. Olsder (1982) , *Dynamic Noncooperative Game Theory*, Academic Press, London/New York.

Bertsekas, D. P. (1987) , *Dynamic Programming: Deterministic and Stochastic Models*, Prentice Hall, Englewood Cliffs, New Jersey.

Blanchard, O. (1979) , "Backward and Forward Solutions for Economies with Rational Expectations", *American Economic Review*, vol.69, pp.114-118.

Blanchard, O. and C. M. Kahn (1980) , "The Solution of Linear Difference Models under Rational Expectations", *Econometrica*, vol.48, no.5, pp.1305-1311.

Kumar, P. R. and P. P. Varaiya (1986) , *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice Hall, Englewood Cliffs, New Jersey.

Lucas, R. (1975) , "An Equilibrium Model of the Business Cycle", *Journal of Political Economy*, vol.83, pp.1113-1144.

Sargent, T. J. and N. Wallace (1975) , "Rational Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply", *Journal of Political Economy*, vol.83, pp.241-254.

Shiller, R. (1978) , "Rational Expectations and the Dynamic Structure of Macroeconomic Models: A Critical Review", *Journal of Monetary Economics*, vol.4, pp.1-44.

Taylor, J. (1977) , "Condition for Unique Solutions in Stochastic Macroeconomic Models with Rational Expectations", *Econometrica*, vol.45, pp.1377-1385.

COLLUSIVE EQUILIBRIA IN STOCHASTIC SEQUENTIAL GAMES WITH LIMIT OF AVERAGE PAYOFFS

ALAIN HAURIE
GERAD, École des Hautes Études Commerciales and École Polytechnique
5255 Decelles, H3T 1V6, Montréal, Canada

BOLESLAW TOLWINSKI
Department of Mathematics
Colorado School of Mines, Golden, Colorado 80401

Abstract This paper deals with the construction of cooperative equilibria for stochastic dynamic games, where the players cannot observe the actions of their opponents. For a particular class of dynamic games with payoffs defined as the limit of average gains one establishes the existence of perfect equibria which are also Pareto-optimal.

1. Introduction

The aim of this paper is to explore the possibility to construct cooperative equilibria in stochastic sequential games of infinite duration with payoffs defined as limit of averages. A stochastic sequential game is a discrete-time dynamic game that involves an element of uncertainty represented by a random noise affecting the state transitions. Sequential games include as a particular case the class of so-called repeated games, which arise when a static game (e.g. a matrix game or a Cournot duopoly game) is played repeatedly over an infinite number of periods.

One of the most interesting features of dynamic game theory is that it allows the study of cooperative or collusive behavior among the agents engaged in the control of a dynamic system, even in the absence of any external mechanism which makes cooperative agreements binding. Recall that the presence of such a "cheating preventing" device is a precondition of cooperation in static games (Luce and Raiffa 1957). Cooperative solutions of dynamic games, on the other hand, can be supported by "cheating-proof" equilibrium strategies, which imply that a player react with a punitive action to any breach of cooperation by his partners. This fact has been first established for deterministic repeated games (see Aumann 1959, Friedman 1977, Rubinstein 1979, 1980, and Radner 1980), and later studied in the more general context of deterministic sequential and differential games in Tolwinski 1982, 1986, Haurie and Tolwinski 1984, 1985, and Tolwinski et al. 1986. The collusive equilibria of dynamic games considered in these works have been obtained under the assumption that a player making his decision at a given instant of time has complete information about his partners' action history. The importance of this assumption stems from the fact that it ensures that each player has the ability to detect any breach of cooperation by other players, and then to react to it in an appropriate manner. The question addressed in the present paper is whether the above assumption can be relaxed, i.e., can a cooperative equilibrium be constructed for a dynamic game, where the players have only incomplete information about other players actions?

The issue of existence of cooperative equilibria in repeated games when the information held by each player about his opponents actions is distorted by random noise has been addressed by Radner 1981 or Rubinstein and Yaari 1983 for the case of average payoff criteria, and Radner 1985 , Porter 1983, Green and Porter 1984 and Fudenberg and Maskin 1986, among others for the case of discounted payoffs. Radner considered the problem of monitoring cooperative agreements in the context of a repeated principal-agent game, and obtained cooperative epsilon equilibria under the assumption that the players maximize their average payoffs over a finite but arbitrarily large number of periods. Radner's approach has been closely related to the idea of sequential tests of power one (Robbins and Siegmund 1974); it takes advantage of the fact that only the changes of policy that are maintained for relatively long periods of time can have any noticeable impact on long-term average payoffs. The changes of this type, on the other hand, can be detected by means of statistical tests. Radner's approach involves a so-called triggering mechanisms: a strategy is then a combination of a cooperative policy, a threat (or punitive) policy, and a switching rule which triggers punitive retaliations. The repeated game structure is not essential for the obtention of such cooperative equilibria as it will be shown in the rest of this paper which establishes a result similar to Radner's in the realm of sequential stochastic games. However an infinite horizon setting is essential for such equilibria to exist, since as shown by Basar 1977 a stochastic sequential game played over a finite time horizon, contrarily to deterministic sequential games typically admits as equilibria only those which correspond to the strictly noncooperative mode of play (viz. the feedback Nash equilibria obtained through the dynamic programming approach). What makes the situation different for infinite horizon stochastic sequential games is the fact that there is always enough time for a player to retaliate if cheating has been detected. When the strategy evaluation criteria defining the players payoffs are the limits of the average transition rewards, then only the long term effects of strategic choices really matter. This fact also facilitates the construction of efficient cooperative equilibria. .

The paper is organized as follows. In Section 2 the definition of the stochastic sequential game is introduced. In Section 3 the definitions of admissible strategy pair, perfect equilibrium, and efficient strategy pair are given. Section 4 is concerned with the extension of Radner's approach for the construction of efficient collusive equilibria, to the class of stochastic sequential games.

2. The stochastic sequential game format

We consider a two-person nonzero-sum discrete-time dynamic stochastic game, also referred to as a *stochastic sequential game* and defined as follows:

Let a dynamic system be described by the state equation

$$x(t+1) = f[x(t),\, u_1(t),\, u_2(t),\, w(t)], \quad t = 0,1,2,\ldots \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the state vector, $u_i(t) \in \mathbb{R}^{m_i}$ is the control variable of Player i, $i = 1,2$, and $\{w(t) \ : \ t = 0,1,2,\ldots\}$ is a purely random sequence of independent identically distributed random variables, with values in $\mathbb{R}^p$; n, m_1, m_2, p are given integers. The function $f : \mathbb{R}^n \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^p \to \mathbb{R}^n$ is given and known by the two players. We assume that each player is able to directly observe the state vector $x(t)$, but can observe neither his opponent's actions nor the realizations of the random variable $w(t)$. As a consequence, Player i selects his control $u_i(t)$ from a given subset U_i of $\mathbb{R}^{m_i}$, on the basis of the information represented by the random sequence

$$\nu(t) = \{x(0), x(1), \ldots, x(t-1), x(t)\}, \quad t = 0, 1, 2, \ldots \tag{2}$$

In other words, a strategy of Player i is defined as a sequence of mappings

$$\gamma_i = \{\gamma_{it} \ : \ t = 0, 1, 2, \ldots\}, \quad i = 1, 2 \tag{3}$$

where γ_{it} associates an element of U_i with every $\nu(t)$. The collection of all strategies of Player i is called Player i's strategy space and it is denoted by Γ_i.

One can view the controlled stochastic system (1) as a family of discrete-time stochastic processes with values in $\mathbb{R}^n$, defined over a measure space (Ω, Σ). The information structure (2) corresponds to an increasing family of σ-fields $\Sigma = \{\Sigma_t \ : \ t = 0, 1, 2, \ldots\}$. A strategy γ_i of Player i is a Σ_t-adapted stochastic process with value in U_i. Associated with any admissible strategy pair $\gamma = (\gamma_1, \gamma_2)$, a probability measure P_γ is defined over (Ω, Σ).

For obvious reasons, a strategy requiring a player to recall the whole sequence $\nu(t)$ for every t, would be of little practical value. Therefore, we consider among the admissible strategies Γ_1 the class of the so-called *Extended Markovian (EM)* strategies. Under the *EM* strategies Player i chooses his control $u_i(t)$ on the basis of an extended state vector $z_i(t) = (x(t), y_i(t))$, where $y_i(t)$ is an auxiliary state variable with values in a given set Y_i and which summarizes the information available to player i concerning the history of the game up to time t. We call $Z = \mathbb{R}^n \times Y_1 \times Y_2$ the set of all possible values for the extended state variable. The evolution over time of the auxiliary state variable $y_i(t)$ is described by an auxiliary state equation of the form

$$\begin{aligned} y_i(0) &= y_i^0, \\ y_i(t+1) &= g_i[x(t), u_1(t), u_2(t), w(t), y_i(t)], \quad t = 0, 1, 2, \ldots \end{aligned} \tag{4}$$

where $g_i : \mathbb{R}^n \times \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \times \mathbb{R}^p \times Y_i \to Y_i$ is a given function, and y_i^0 is a given initial value.

A stationary *EM* strategy γ_i in Γ_i is such that for every t, γ_{it} is a function of the extended state alone, i.e., γ_{it} does not explicitly depend on t. In such a case, the symbol γ_i will be used to denote γ_{it}, i.e., Player i's decision rule at stage t, as well as his strategy, i.e., the whole infinite sequence of those rules Also one should notice that an *EM* strategy implies a specific information structure associated with the auxiliary state equation (4).

Remark 1. *The choice of the auxiliary variable $y_i(t)$ and state equation (4) is part of the design of a strategy by Player i. In some strategy designs the auxiliary state variable $y_i(t)$ can be used as an indicator of the mood of play. $y_i(t) = 1$ indicates that a cooperative mood of play prevails whereas $y_i(t) = 0$ indicates that a noncooperative mood of play is adopted. In section 4 cooperative equilibria will be obtained in this class of EM strategies.*

Remark 2. *The so-called stationary feedback strategies for which the extended state variable at any time t reduces to $z(t) = x(t)$ constitute a particular subclass of EM strategies. This class of strategies has been the object of most of the attention devoted to the theory of stochastic sequential games (Sobel 1971, Whitt 1980).*

3 Strategy evaluation criteria (payoff functionals) and equilibria

Let $h_i \ : \ U_1 \times U_2 \times \mathbb{R}^n \rightarrow \mathbb{R}$ represent the transition reward function of Player i, $i = 1,2$. We say that the strategy γ is admissible if the following payoff functionals are well defined for any initial extended state $z(0)$

$$J_i[z(0);\ \gamma] = \lim_{T\rightarrow\infty} (1/T)\, E_\gamma \left\{ \sum_{t=0}^{T-1} h_i[x(t),\ u_1(t), u_2(t)] \right\} \tag{5}$$

In the above formula, x, , u_1, u_2 are stochastic processes associated with the strategy pair γ and the expectation is taken with respect to the probability measure induced by the strategy pair γ, where

$$\gamma = (\gamma_1, \gamma_2) = \{\gamma_t = (\gamma_{1t}, \gamma_{2t})\ :\ t = 0, 1, 2, \ldots\} \tag{6}$$

The expression (6) defines the so-called *limit of average criteria* for strategy evaluation in this infinite horizon stochastic sequential game.

Remark 3. *The use of a limit of average criterion implies that the player is only concerned by the lasting effects of his strategic choices. More precisely any effect which appears in a finite number of transitions will become negligible.*

Remark 4. *With the limit of average criterion it often happens that the payoffs associated with a stationary feedback strategies do not depend on the initial state. This ergodicity property will be exploited in the construction of cooperative equilibria.*

Definition 1. *An EM strategy pair $\gamma^* = (\gamma_1^*, \gamma_2^*)$, associated with the extended state $z(t)$ and the auxiliary state equation (4), is a (perfect) equilibrium if it is admissible and at any initial extended state $z(0) \in \mathbb{R}^n$ the following holds*

$$J_1[z(0); \gamma^*] \geq J_1[z(0); \gamma_1, \gamma_2^*] \tag{7}$$

for all $\gamma_1 \in \Gamma_1$ such that (γ_1, γ_2^*) is admissible and

$$J_2[z(0); \gamma^*] \geq J_2[z(0); \gamma_1^*, \gamma_2] \tag{8}$$

for all $\gamma_2 \in \Gamma_2$ such that (γ_1^*, γ_2) is admissible.

Definition 2. *An EM strategy pair $\gamma^* = (\gamma_1^*, \gamma_2^*)$ is efficient if it is admissible and at any initial extended state $z(0) \in \mathbb{R}^n$ the following holds for any γ admissible*

$$J_i[z(0); \gamma] \geq J_i[z(0); \gamma^*] \ \ i = 1,2 \Rightarrow J_i[z(0); \gamma] = J_i[z(0); \gamma^*] \ \ i = 1,2 \tag{9}$$

4. Efficient Cooperative Equilibria

In this section we extend to a sequential game format the approach initially proposed by Radner 1981 in the realm of repeated games, for the construction of efficient perfect equilibria. The following assumptions are assumed to hold:

(A1) The game has a stationary feedback equilibrium $\mu = (\mu_1, \mu_2)$ generating payoffs whose values are independent of the initial state of the system, i.e.,

$$J_i(x_0; \mu) = V_i^N = \text{const. for every } x_0 \in \mathbb{R}^n, \quad i = 1,2 \tag{10}$$

In addition, there exists an efficient stationary feedback strategy $\eta = (\eta_1, \eta_2)$ such that

$$J_i(x_0; \eta) = V_i^C = \text{const. for every } x_0 \in \mathbb{R}^n, \quad i = 1,2 \tag{11}$$

and

$$V_i^C > V_i^N \text{fori} = 1,2. \tag{12}$$

(A2) There exist numbers M_1 and M_2 such that

$$| h_i[x, \eta_i(x)] | \leq M_1 (1+ \| x \|), \quad i = 1,2 \tag{13}$$

$$\| f[x, \eta(x), w] \| \leq M_2 (1+ \| w \|) \tag{14}$$

for every $x \in X$.

(A3) The components of the random vector w have finite expected values and variances.

The random variables defined below will serve as statistics for monitoring adherence to cooperative policies η during the play. Let

$$x_\eta(t) = f[x(t-1), \eta(x(t-1)), w(t-1)] \tag{15}$$

and

$$e_i(t) = E\,\{h_i[x_\eta(t), \eta(x_\eta(t))|x(t-1)]\}, \quad i = 1,2. \tag{16}$$

Consider the stochastic processes

$$\Psi_1(t) = h_1[x(t), \eta_1(x(t)), u_2(t))] - e_1(t)$$

and

$$\Psi_2(t) = h_2[x(t), u_1(t)), \eta_2(x(t))] - e_2(t), \tag{17}$$

and define

$$S_i(t) = \sum_{s=0}^{t} \Psi_i(s), \quad t = 1,2,\ldots, \quad i = 1,2 \tag{18}$$

We shall denote by $\hat{\Psi}_i(t)$ and $\hat{S}_i(t)$ the values of $\Psi_i(t)$ and $S_i(t)$ respectively, corresponding to the case when $u_j(t) = \eta_j(x(t))$ for all t and $j \neq i$.

Remark 5. *The random variable $x_\eta(t)$ defined in (15) is the state that would result from the use of the strategy pair η at period $t-1$ and at state $x(t-1)$ (recall that η is a feedback strategy pair). The conditional expected value $e_i(t)$, given the observed state $x(t-1)$ defined in (16) can thus be computed by Player i at each period t. The stochastic process $\Psi_i(t)$ defined in (17) is thus based on a comparison between the conditional expected transition reward when the cooperative feedback strategy pair prevails and the actual realization of this reward. This will provide the information basis permitting Player i to detect cheating by his opponent provided that he can observe his own transition rewards.*

Lemma 1. *Under (A2) and (A3), $\hat{S}_i(t)/t$ converges to zero almost surely.*

Proof: This result is a direct consequence of the generalized Strong Law of Large Numbers (Feller 1971, page 243, Theorem 3), provided that $E\{\hat{\Psi}_i(t)^2\}$ can be shown to be bounded for all $t \in \{1,2,\ldots\}$

To see that the latter is true, observe that

$$E\{\hat{\Psi}_i(t)^2\} = E\{h_i[x_\eta(t),\, \eta_i(x_\eta(t))]^2\} - e_i(t)^2 \tag{19}$$

In view of (A2) one has

$$\begin{aligned} |\, h_i[x(t),\, \eta_i(x(t))]\, | &\le M_1[1+ \|\, x(t)\, \|] \\ &= M_1[1+ \|\, f[x(t-1),\, \eta(x(t-1)), w(t-1)]\, \|] \\ &\le M_1[1 + M_2(1+ \|\, w(t-1)\, \|) \\ &= M[1+ \|\, w(t-1)\, \|] \end{aligned} \tag{20}$$

where M is a constant depending on M_1 and M_2. Hence,

$$|\, e_i(t)\, |=|\, E\{h_i\}\, | \le M[1 + E\{\|\, w(t-1)\, \|\}] \tag{21}$$

and

$$E\{h_i^2\} \le M^2[1 + 2E\{\|\, w(t-1)\, \|\} + E\{\|\, w(t-1)\, \|^2\}] \tag{22}$$

Therefore, (A3) implies that the variances of $\hat{\Psi}_i(t)$ are bounded for all t. •

We now proceed to the construction of an efficient cooperative equilibrium defined by an *EM* strategy pair. Let $\{b_i(t)\}$, $i = 1,2$ be two sequences of positive numbers such that $b_i(t)$ tends to infinity, and $b_i(t)/t$ converges to zero when t approaches infinity. We define the auxiliary state variables, $y_i(t)$, $i = 1,2$, with value in $Y_i = \{0,1\}$ and the following dynamics

$$\begin{aligned} y_i(0) &= 1 \\ y_i(t) &= \begin{cases} 1 & \text{if } y_i(t-1) = 1 \text{ and } S_i(t)/t \ge -b_i(t)/t \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{23}$$

we define an *EM* strategy pair, $\gamma = (\gamma_1, \gamma_2)$, as follows

$$\gamma_i(x(t),\, y_1(t),\, y_2(t)) = \begin{cases} \eta_i(x(t)) & \text{if } y_j(t) = 1 \text{ for } j = 1,2 \\ \mu_i(x(t)) & \text{otherwise} \end{cases} \tag{24}$$

Remark 6. *The dynamics for the information variable $y_i(t)$ given by Eq. (23) does not seem at first sight to be of the general form described in Eq. (4). However it would be easy and straightforward to obtain the form given in (4) by noticing that the variables $S_i(t)$ satisfy the following state equation*

$$S_i(t+1) = S_i(t) + \Psi_i(t) \quad ,i = 1,2.$$

Remark 7. *The strategy γ_i is such that Player i begins the play in a cooperative (Pareto efficient) mood of play; however if some "cheating" is detected at period t through the mechanism defined by (17), (18) and (23), then Player i switches to a purely noncooperative mood of play (feedback Nash equilibrium) and maintains it forever.*

Proposition 1. *There exists a sequence $\{b(t)\}$ such that the strategy pair γ defined by (24)-(25) constitutes a perfect equilibrium, and the payoffs generated by γ coincide with V_i^C, i.e., the equilibrium γ is also efficient.*

Proof: If $y_1(t)$ or $y_2(t)$ is zero, then both players use policies (μ_1, μ_2), which constitute an equilibrium by (A1). Now, consider the cases when $y_1(t) = y_2(t) = 1$. Since the policies (η_1, η_2) are Pareto-optimal, any deviation from η_i which leads to an increase in the payoff of player i, must at the same time cause a decrease in the payoff of the other player. Suppose that Player 1 has unilaterally changed η_1 for a policy ζ_1 which generates a sequence of one step payoffs which satisfy:

$$E_{(\zeta_1,\eta_2)}\{h_1[x(t),\ u_1(t)], u_2(t)\} = E_\eta\{h_1[x_\eta(t),\ \eta_1(x_\eta(t)), \eta_2(x_\eta(t))]\} + \delta_1(t) \tag{25}$$

If the above change of policy is to have any effect on the overall payoff of player 1, one must have

$$\lim_{t\to\infty}(1/T)\sum_{t=0}^{T-1}\delta_1(t) \geq \delta_1 > 0 \tag{26}$$

for some number δ_1, because otherwise

$$\begin{aligned}\lim_{T\to\infty}(1/T)\,E_{(\zeta_1,\eta_2)}\left\{\sum_{t=0}^{T-1} h_1[x(t),\ u_1(t), u_2(t)]\right\} &= \lim_{T\to\infty}(1/T)\,E\left\{\sum_{t=0}^{T-1} h_1[x_\eta(t),\ \eta(x_\eta(t))]\right\}\\ &\quad + \lim_{T\to\infty}(1/T)\sum_{t=0}^{T-1}\delta_1(t)\\ &\leq V_1^C + \lim_{T\to\infty}(1/T)\sum_{t=0}^{T-1}\delta_1 \leq V_1^C\end{aligned} \tag{27}$$

In the case when (26) holds, Player 2 will receive a payoff of the form

$$E_{(\zeta_1,\eta_2)}\{h_2[x(t), u_1(t), u_2(t)]\} = E_\eta\{h_2[x_\eta(t),\ \eta(x_\eta(t))]\} + \delta_2(t) \tag{28}$$

where

$$\lim_{T\to\infty}(1/T)\sum_{t=0}^{T-1}\delta_2(t) \leq \delta_2 < 0 \tag{29}$$

for some number δ_2. Hence, in view of Lemma 1,

$$\lim_{t\to\infty} S_2(t)/t \leq \lim_{t\to\infty}\hat{S}_2(t)/t + \delta_2 \leq \delta_2 \quad \text{a.s.} \tag{30}$$

Since $b_2(t)/t$ approaches zero when t tends to infinity, (30) implies that $y_2(t)$ will eventually become zero almost surely. In other words, any deviation from η_1 which could lead to an increase in Player 1's payoff will almost surely be detected by Player 2, who will then switch to μ_2. The best response of Player 1 in such case will be to switch to μ_1, and his payoff resulting from this turn of events will be V_1^N. To complete the proof, it suffices to show that, provided that the sequences $\{b_i(t)\}$, $i = 1,2$ be conveniently defined, if Player 1 stays with η_1. then he will receive V_1^C, which by assumption is greater than V_1^N.

By Lemma 1 we know that for every integer m, there exists a number T_m such that

$$P\{S_i(t)/t \geq -1/m, \quad \text{for} \quad t \geq T_m\} = 1, \quad i = 1,2 \tag{31}$$

Hence, if we define the sequences $\{b_i(t)\}$ $i = 1,2$ in such a way that they satisfy

$$b_i(t) = \infty \quad \text{for} \quad 1 \leq t < T_1, \quad b_i(t) = t/m \quad \text{for} \quad T_m \leq t < T_{m+1}, \quad m = 1,2,\ldots \tag{32}$$

then $S_i(t)/t \geq -b_i(t)/t$ almost surely for all t and $i = 1,2$. Therefore, assuming that Player 2 uses γ_2, if Player 1 does not deviate from η_1, then $y_1(t)$ and $y_2(t)$ will almost surely remain equal to one for all t, which means that Player 1's expected payoff corresponding to the use of strategies (24)-(25) will be equal to V_1^C.

Since the same argument as the one given above applies to the analysis of consequences of deviations from γ_2 by Player 2, we have shown that the strategy pair given by the expression (25) with b_i's satisfying (32) is in fact a perfect equilibrium and that it generates Pareto-optimal payoffs. Thus, the proof has been completed. •

5. Conclusion

We have shown that a class of sequential games with limit of averages payoffs admit efficient collusive equilibria in the class of extended markovian strategies. This result completes and extends Radner's theory.

REFERENCES

Aumann, R.J., 1959, "cooperative n-person games", in: A. W. Tucker and R. D. Luce (Eds.), *Contribution to the Theory of Games*, Vol. IV, 287–324, Princeton.

Basar, T., 1977, "Informationally nonunique equilibrium solutions in differential games", *SIAM J. Control Optimiz.*, Vol. 15, 636–660.

Breton M., J. A. Filar, A. Haurie and T. A. Schulz, 1986, "On the computation of equilibria in discounted stochastic dynamic games", in: T. Basar (Ed.) *Dynamic Games and Applications in Economics*, Springer Verlag.

Feller, W., 1971, *An Introduction to Probability Theory and its Applications*, Vol. II, Second Edition, Wiley.

Friedman, J. W., 1977, *Oligopoly and the Theory of Games*, North Holland.

Fudenberg, D. and E. Maskin, 1986, "The Folk theorem in repeated games with discounting and with incomplete information", *Econometrica*, Vol. 54, 533–554.

Green, E. J. and R. H. Porter, 1984, "Noncooperative collusion under imperfect price information", *Econometrica*, Vol. 52, 87–100.

Haurie, A. and B. Tolwinski, 1984, "Acceptable equilibria in dynamic bargaining games", *Large Scale Systems*, Vol. 6, 73–89.

Haurie, A. and B. Tolwinski, 1985, "Definition and properties of cooperative equilibria in a two-player game of infinite duration", *J. Optimiz. Theory Appl.*, Vol. 46, 525–533.

Luce, R. D. and H. Raiffa, 1957, *Games and Decisions*, Wiley.

Porter, R. H., 1983, "Optimal cartel trigger price strategies", *J. Econ. Theory*, Vol. 29, 313–338.

Radner, R., 1980, "Collusive behavior in noncooperative epsilon equilibria of oligopolies with long but finite lives", *J. Econ. Theory*, Vol. 22, 136–154.

Radner, R., 1981, "Monitoring cooperative agreements in a repeated principal-agent relationship", *Econometrica*, Vol. 49, 1127–1148.

Radner, R., 1985, "Repeated principal-agent games with discounting", *Econometrica*, Vol. 53, 1173–1198.

Robbins, H. and D. Siegmund, 1974, "The expected sample size of some tests of power one", *The Annals of Statistics*, Vol. 2, 415–436.

Rubinstein, A., 1979, "Equilibrium in supergames with the overtaking criterion", *J. Econ. Theory*, Vol. 21, 1–9.

Rubinstein, A., 1980, "Strong perfect equilibrium in supergames", *Int. J. Game Theory*, Vol. 9, 1–12.

Tolwinski, B., 1982,"A concept of cooperative equilibrium for dynamic games", *Automatica*, Vol.18, 431–447.

Tolwinski, B., 1987, "A renegotiation-proof solution for a price setting duopoly", "1987 Optimization Days", Montréal, Canada.

Tolwinski, B., A. Haurie and G. Leitmann, 1986, "Cooperative equilibria in differential games", *J. Math. Anal. Appl.*, Vol. 119, 182–202.

Selten, R., 1975, "Reexamination of the perfectness concept for equilibrium points in extensive games", *Int. J. Game Theory*, Vol. 4, 25–55.

Whitt, W., 1980, "Representation and approximation of noncooperative sequential games", *SIAM J. Control Optimiz.*, Vol. 18, 33–48.

OPTIMAL BAYESIAN CONTROL OF
A NONLINEAR REGRESSION PROCESS
WITH UNKNOWN PARAMETERS

by

Nicholas M. Kiefer
Department of Economics
Cornell University
Ithaca, New York 14853

and

Yaw Nyarko
Department of Economics
Brown University
Providence, RI 02912

1. Introduction

Economic Agents operating in uncertain, stochastic environments can face a tradeoff between current period expected reward and accumulation of information of uncertain value. For example, a firm producing to meet uncertain demand might produce at the expected current reward maximizing output, based on his current beliefs about the form of the demand curve, or it might choose to experiment by varying outpuc, thus taking short term losses in order to sharpen beliefs about the form of the demand curve. A parametric representation of the agent's problem is made by considering the utility function $u(x,y)$ and the conditional density $f(y|x,\theta)$. Here the random variable y is what the agent is trying to control (e.g., current period profits) and x is the control variable. The parameters θ of the conditional density of y given x are unknown, but the agent has opinions about θ given by a distribution μ. The agent attempts to minimize the present discounted value of the stream of expected losses, $E\Sigma\delta^t u(x_t,y_t)$, where the expectation is taken with respect to current beliefs. The problem is complicated by the fact that beliefs are updated from period to period using Bayes Rule; consequently current period actions can be expected to influence future period beliefs. This introduces stochastic dynamics into the model.

This paper considers the problem in the case in which the density $f(y|x,\theta)$ is a location family. In this case the model can be written $y - g(x,\beta) + \epsilon$, where ϵ is an i.i.d. random variable whose distribution may involve unknown parameters. When $g(x,\beta) - x'\beta$ the problem is one of controlling a linear regression process with unknown parameters over an infinite horizon. Many approximate control rules for this problem have been proposed, for example sequential least-squares estimation combined with one-period optimization conditioning on the current estimates. The analogous policy for the nonlinear model is clear. In practice several policies can work "well," though it is possible to compose examples in which the policy mentioned, for example, is easily improved. From an economic modelling point of view, however, we are interested in the *optimal* policy, and in the consequences for convergence of beliefs and policies of following the optimal policy. Will it be optimal for an agent to learn the parameters (and thus converge to "rational expectations")?

This paper gives general conditions under which the sequence of beliefs converges to a limit and the sequence of optimal policies converges to a limit. Under further conditions the limit policy is the optimal one-period policy for limit beliefs. Conditions under which the limit belief is point mass at true parameter values, corresponding to consistent parameter estimates are more stringent and are still under investigation.

Least-squares control rules in the linear regression model have been widely discussed and studied analytically by Taylor (1974) and Jordan (1985) and experimentally by Anderson and Taylor (1976). Improvements using a Bayesian approach were suggested by Zellner (1971) and studied by Harkema (1975). The optimal policy in the linear regression case has been studied by Kiefer and Nyarko (1987), who obtain results on convergence of beliefs and policies. convergence in a different class of models has been studied by Easley and Kiefer (1986). Results on optimal learning while controlling a stochastic process are collected along with an example in Kiefer (1988).

2. The Decision Problem: Uncertainty, Policies and Rewards

In this section we sketch the general framework we wish to study.

Let Ω' be a complete and separable metric space, let $\mathcal{F}'$ be its Borel field, and $(\Omega', \mathcal{F}', P')$ a probability space. Define the stochastic process $\{\epsilon_t\}^\infty$ on $(\Omega', \mathcal{F}', P')$. The ϵ_t are assumed to be independent and identically distributed, with the common marginal distribution $p(\epsilon_t|\xi)$ depending on some parameter, ξ in R^h, which is unknown to the agent. We assume that the set of probability measures, $\{p(\cdot|\xi)$, is continuous in the parameter ξ (in the weak topology of measures); and that for any ξ, $\int \epsilon \, p(d\epsilon|\xi) = 0$. Let $\tilde{X}$, the *action space*, be a compact subset of R^k. Define $\Theta = R^m \times R^h$ to be the parameter space. If the "true parameter" is $\theta = (\beta,\xi) \in \Theta$, and the agent chooses an action $x_t \in \tilde{X}$ at date t, then the agent observes y_t, where,

$$y_t = g(x_t,\beta) + \epsilon_t \tag{2.1}$$

and ϵ is chosen according to $p(\cdot|\xi)$. The function g is assumed measurable;

further restrictions are introduced implicitly through assumptions on the updating equation (2.2) and the reward function (2.3).

One example is the simple linear regression model with unknown slope and intercept and with the ϵ_t independent draws from the normal distribution with mean zero and variance σ^2. In that example Ω' is R^∞, $\mathcal{F}$ is the collection of Borel sets on R^∞, and P' is the infinite product of independent univariate normal distributions with means zero and common variance σ^2. The parameter ξ is the variance of ϵ, σ^2. The action space $\bar{X}$ is a closed interval in R^1. The parameter $\beta \in R^2$ consists of the slope and intercept of the regression. The space Θ is $R^2 \times R^1_+$.

Let $\mathcal{J}$ be the Borel field of Θ, and let $P(\Theta)$ be the set of all probability measures on $(\Theta, \mathcal{J})$. Endow $P(\Theta)$ with its weak topology, and note that $P(\Theta)$ is then a complete and separable metric space (see e.g., Parthasarathy (1967, Ch. II, Theorems 6.2 and 6.5)). Let $\mu_0 \in P(\Theta)$ be the prior probability on the parameter space, with finite first moment.

The agent is assumed to use Bayes rules to update the prior probability at each date after any observation of (x_t, y_t). For example, in the initial period, date 1, the prior distribution is updated after the agent chooses an action x_1, and observes the value of y_1. The updated prior, i.e., the posterior, is then $\mu_1 = \Gamma(x_1, y_1, \mu_0)$, where $\Gamma:\bar{X} \times R^1 \times P(\Theta) \to P(\Theta)$ represents the Bayes rule operator. If the prior, μ_0, has a density function, then the posterior may be easily computed. In general, the Bayes rule operator may be defined by appealing to the existence of certain conditional probabilities, although some care is needed (see Diaconis and Freedman (1986)). Under some conditions the operator Γ is continuous in its arguments, and we assume this throughout. Any $\{x_t, y_t\}$ process will therefore result in a posterior process, $\{\mu_t\}$, where for all $t = 1,2,\ldots,$

$$\mu_t = \Gamma(x_t, y_t, \mu_{t-1}) \tag{2.2}$$

Let $\hat{H}_n = P(\Theta) \times \prod_{i=1}^{n-1} [\bar{X} \times R^1 \times P(\Theta)]$. A <u>partial history</u>, h_n, at date n is any element $h_n = (\mu_0,(x_1, y_1, \mu_1)\ldots,(x_{n-1}, y_{n-1},\mu_{n-1})) \in \hat{H}$; h_n is said to be admissible if (2.2) holds for all $t = 1,2,\ldots, n-1$. Let $\bar{H}_n$ be the subset of $\hat{H}_n$ consisting of all admissible partial histories at date n. A <u>policy</u> is a sequence $\pi = \{\pi_t\}_{t=1}^\infty$, where for each $t \geq 1$, the policy function $\pi_t:H_t \to \bar{X}$ specifies the date t action $x_t = x_t(h_t)$, as a Borel function of the partial history, h_t in

H_t, at that date. A policy function is stationary if $\pi_t(h_t) = g(\mu_t)$ for each t, where the function $g(\cdot)$ maps $P(\Theta)$ into $\hat{X}$.

Define $(\Omega, \mathcal{F}, P) = (\Theta, \mathcal{F}, \mu_0) \times (\Omega', \mathcal{F}', P')$. Any policy, π, then generates a sequence of random variables $\{(x_t(\omega), y_t(\omega), \mu_t(\omega)\}_{t=1}^{\infty}$ on $(\Omega, \mathcal{F}, P)$ as described above, using (2.1) and (2.2). See Kiefer and Nyarko (1987) for technical details.

For any $n = 1,2,\ldots,$ let $\mathcal{F}_n$ be the sub-field of $\mathcal{F}$, generated by the random variables (h_n, x_n). Notice that x_n is $\mathcal{F}_n$ -measurable but y_n and μ_n are not $\mathcal{F}_n$ - measurable. Next define $\mathcal{F}_\infty = \vee_{n=0}^{\infty}\mathcal{F}_n$.

Let $u:\hat{X} \times R^1 \to R^1$ be the utility function, so $u(x_t, y_t)$ is the utility to the agent when action x_t is chosen at date t and the observation y_t is made. The reward function $r:\hat{X} \times P(\Theta) \to R^1$, is defined by

$$r(x_t,\mu_{t-1}) = \int_\Theta\int_R u(x_t, y_t)p(d\epsilon_t|\xi)\mu_{t-1}(d\theta) \tag{2.3}$$

The inner integration marginalizes with respect to ϵ, given the parameter ξ, the outer integration is with respect to parameters. Assume that the reward function is uniformly bounded, continuously, and concave in x for given μ. Note that this assumption restricts $g(\cdot,\cdot)$, $U(\cdot,\cdot)$ and $p(\cdot|\cdot)$.

Let δ in $[0,1)$ be the discount factor. Any policy π generates a sum of expected discounted rewards equal to

$$V_\pi(\mu_0) = \int \sum_{t=1}^{\infty} \delta^{t-1} r(x_t(\omega), \mu_{t-1}(\omega))P(d\omega) \tag{2.4}$$

where the (x_t, μ_t) processes are those obtained using the policy π. A policy π^* is said to be an optimal policy if for all policies π and all priors μ_0 in

in $P(\Theta)$, $V_{\pi^*}(\mu_0) \geq V\ (\mu_0)$. Even though the optimal policy, π^* (when it exists) may not be unique, the value function $V(\mu_0) = V_{\pi^*}(\mu_0)$ is always well-defined.

3. Existence of a Stationary Optimal Policy

Straightforward dynamic programming arguments can be used to show that stationary optimal policies exist and the value function is continuous.

Theorem 3.1: A stationary optimal policy $g:P(\Theta) \to \hat{X}$ exists. The value function, V, is continuous on $P(\Theta)$, and the following functional equation holds:

$$V(\mu) = \max \{r(x, \mu) + \delta\int V(\tilde{\mu})p(d\varepsilon|\xi)\mu(d\theta)\} \qquad (3.1)$$

where $\tilde{\mu} = \Gamma(x, y, \mu)$ and $y = g(x, \beta) + \varepsilon$, and where the integral is taken over $R^1 \times \Theta$.

<u>Proof</u>: Let $S = \{f:P(\Theta) \to R \mid f$ is continuous and bounded$\}$.

Define $T:S \to S$ by

$$Tw(\mu) = \max_{x\varepsilon\bar{X}} \{r(x,\tilde{\mu}) + \delta\int V(\mu)p(d\varepsilon|\phi)\mu(d\theta)\} \qquad (3.2)$$

One can easily show that for $w\varepsilon S$, $Tw\varepsilon S$; and that T is a contraction mapping. Hence there exists a $v\varepsilon S$ such that $v=Tv$. Replacing w with v in (3.2) then results in (3.1); and since $v\varepsilon S$, v is continuous. Finally, it is immediate that the solution to the maximization exercise in (3.2) (replacing w with v) results in a stationary optimal policy function (see Blackwell (1965) or Maitra (1968) for the details of the above arguments).

4. Convergence of the Process $\{\mu_t\}$.

In this section we prove that the posterior process converges for P-a.e ω in Ω, to a well-defined probability measure (with the convergence taking place in a weak topology).

Note that for any Borel subset, D, of the parameter space Θ, if we suppress the ω's and let, for some fixed ω, $\mu_t(D)$ represent the mass that measure $\mu_t(\omega)$ assigns to the set D, then

$$\mu_t(D) = E[1_{\{\theta\varepsilon D\}}|\mathcal{F}_t] \qquad (4.1)$$

Define a measure μ_∞ on Θ by setting, for each Borel set D in Θ,

$$\mu_\infty(D) = E[1_{\{\theta\varepsilon D\}}|\mathcal{F}_\infty] \qquad (4.2)$$

The measure μ_∞ is the limiting posterior distribution and is indeed a well-defined probability measure.

<u>Theorem 4.1</u>. The posterior process $\{\mu_t\}$ converges, for P-a.e. ω in Ω, in the weak topology, to the probability measure μ_∞.

<u>Summary of Proof</u>: Use (4.1) above to show that for any Borel set D in Θ, $\mu_t(D)$ is a Martingale measure, establish that the sequence of probability

measures, $\mu_t(\omega)$, for fixed ω, is tight using the assumption that the first moment of μ_∞ is finite, then apply Prohorov's Theorem (e.g., Billingsley (1968, Theorem 6.1)) to deduce that μ_∞ is a probability measure.

Note that this result on convergence of beliefs is quite different from the standard consistency result looked for in econometrics. The Martingale Convergence Theorem allows us to establish convergence, but the limit measure μ_∞ is a random variable, in the sense that it depends on the particular sequence of shocks realized. In a standard estimation problem, the limit result is that beliefs converge and the limit belief is independent of sample paths, and the limit belief is correct in the sense that μ_∞ assigns point mass to the true parameter value. Standard results do not hold here because along any sample path for which beliefs converge, the sequence of actions $\{x_t\}$ may also be converging. But if actions converge too rapidly, they may not generate enough information to identify all the unknown parameters. One can construct examples in related problems in which this phenomenon occurs (see e.g., Kiefer (1988)).

5. Optimization and Limit Beliefs and Actions

In Theorem 4.1, convergence of beliefs was established for an arbitrary $\{x_t\}$ sequence (i.e., without taking into account the underlying maximization problem). In this section we ask what action (or actions) $\dot{x}$ corresponds to the limiting beliefs μ_∞.

Theorem 5.1 establishes that the limit action is the action which maximizes single period reward for limit beliefs.

Theorem 5.1: The limit action $\dot{x} = \lim_{t\to\infty} x$ exists, is unique for given μ) and maximizes the one-period reward, $r(x,\mu_\infty)$, for limit beliefs μ_∞.

Proof of Theorem 5.1: Recall from Theorem 4.1 that $\lim_{t\to\infty} \mu_t = \mu_\infty$ exists for all sample paths. The sequence $\{x_t\}$ and $\{\mu_t\}$ satisfies for each t (simultaneously, a.e.) the functional equation

$$V(\mu_t) = r(x_t,\mu_t) + \delta\int V(\Gamma(x_t,y_t,\mu_t))p(d\epsilon|\xi)\mu_t(d\theta). \qquad (5.1)$$

Taking limits along any convergent subsequence gives

$$V(\mu_\infty) = r(\bar{x},\mu_\infty) + \delta\int V(\Gamma(\bar{x},y,\mu_\infty))p(d\varepsilon|\xi)\mu_\infty(d\theta)$$

where $\bar{x}$ is a limit point of the (x_t) sequence. (In taking the limits one uses the fact that V is bounded and the integral in (5.1) is $E[V(\mu_t)|\mathcal{F}_{t-1}]$ to apply Chung (1974, Theorem 9.4.8).) However, from convergence of beliefs $(\bar{x},y)$ yields no information so $\Gamma(\bar{x},y,\mu_\infty) = \mu_\infty$, and (5.1) becomes $V(\mu_\infty) = r(\bar{x},\mu_\infty) + \delta V(\mu_\infty)$.

Now we show that $\bar{x}$ solves the problem

$$\max_{x\epsilon\bar{X}} r(x,\mu_\infty) \tag{5.2}$$

Suppose on the contrary that there is an $\hat{x}\epsilon\bar{X}$ such that $r(\hat{x},\mu\ ) > r(\bar{x},\mu\ )$. Then by the functional equation

$$V(\mu_\infty) \geq r(\hat{x},\mu_\infty) + \delta\int V(\Gamma(\hat{x},\hat{y},\mu_\infty))p(d\varepsilon|\Theta)\mu_\infty(d\theta). \tag{5.3}$$

But by Blackwell's Theorem (see e.g., Kihlstrom (1984, Lemma 1, p. 18)), since the experiment "observe $(\hat{x},\hat{y})$" is trivially sufficient for the experiment "make no observations," we obtain,

$$\int V(\Gamma(x,y,\mu_\infty))p(d\varepsilon|\phi)\mu_\infty(d\theta) \geq V(\mu\ ) \tag{5.4}$$

Hence, from (5.3) and (5.4) $V(\mu\ ) > r(\bar{x},\mu\ ) + \delta V(\mu\ )$, which is a contradiction. So $\bar{x}$ solves problem (5.2); that is, $\bar{x}$ maximizes the one-period reward $r(x,\mu)$ for limit beliefs, μ . Since $r(\cdot,\mu_\infty)$ is strictly concave in x, $\bar{x}$ must be unique.

6. Conclusion

We have considered the decision problem facing an agent controlling a nonlinear regression process when parameters in the mean function and in the error distribution are unknown. The agent faces a tradeoff between accumulating information by varying the values of the regressors and accumulating one-period reward by following the one-period expected reward maximizing policy. We show that the problem can be brought into the dynamic programming framework and that the value function satisfies the usual functional equation. The sequence of beliefs about the unknown parameters is shown to converge almost surely. Further, the optimal action process converges to the one-period optimal action under limit beliefs.

7. Acknowledgements

This research is supported in part by the National Science Foundation.

REFERENCES

Anderson, T.W. and J. Taylor, (1976), "some Experimental Results on and Statistical Properties of Least Squares Estimates in Control Problems," Econometrica, 44:1289-1302.

Billingsley, P., (1968), Convergence of Probability Measures, Wiley, New York.

Blackwell, D., (1965), "Discounted Dynamic Programming," Annals of Mathematical Statistics, 36, pp. 2226-235.

Chung, K.L., (1974), A Course in Probability Theory, 2nd edition, Academic Press, New York.

Diaconis, P. and D. Freedman, (1986), "On The Consistency Of Bayes Estimates," Annals of Statistics, 14, 1-26 (discussion and rejoinder 26-27).

Easley, D. and N.M. Kiefer, (1986), "Controlling a Stochastic Process with Unknown Parameters," Cornell University working paper, forthcoming in Econometrica.

Harkema, R., (1975), "An Analytical Comparison of Certainty Equivalence and Sequential Updating," JASA, 70, 348-350.

Kiefer, N.M. and Y. Nyarko, "Control of a Linear Regression Process with Unknown Parameters" in W. Barnett, E. Berndt and H. White (eds.), Dynamic Econometric Modelling, New York: Cambridge University Press, 1987.

Kiefer, N.M., "Optimal Collection of Information by Partially Informed Agents," Cornell working paper, 1988.

Kihlstrom, R.E., (1984), "A 'Bayesian' Exposition of Blackwell's Theorem on the Comparison of Experiments," in Bayesian Models in Economic Theory, eds. M. Boyer and R.E. Kihlstrom, Elsevier Science Publishers B.V.

Jordan, J.S., (1985), "The Strong Consistency of the Least Squares Control Rule and Parameter Estimates," manuscript.

Maitra, A., (1968), "Discounted Dynamic Programming in Compact Metric Spaces," Sankhya, Ser A, 30, pp. 211-216.

Parthasarathy, K., (1967), Probability Measures on Metric Spaces, Academic Press, New York.

Taylor, J.B., (1974), "Asymptotic Properties of Multiperiod Control Rules in the Linear Regression Model," International Economic Review, 15, 472-484.

Zellner, A., (1981), An Introduction to Bayesian Inference in Econometrics, Wiley: New York.

INFORMATION AND DECISION IN OPTIMAL INVENTORY PROCESSES

Toshio Odanaka
Tokyo Metropolitan Institute of Technology
6-6, Asahigaoka, Hino-city, Tokyo, 191, Japan

ABSTRACT

According to developments in management information systems, more investigation is required to adapt the fundamental features that American management information systems have to the Japanese technical climate. One important problem is to decide the kind and the accuracy of management information systems. If complete information is desired regarding a system in each stage of control, some time and cost will be entailed. Otherwise, if incomplete information make a decision quickly, we must put up with using a probability that control a non-optimum system. We have not the complete accuracy for the information and the decision both. This is analogous to Heisenberg's uncertainty principle. In this paper, we discuss the relation between the information and the decision in optimal inventory processes in this viewpoint.

INTRODUCTION

According to management information system development, more investigation is required before adapting the funiamental features of the American management information system to the Japanese technical climate. One important problem is to decide the types and the accuracy of such a management information system. If complete information is desired regarding a system in each stage of control, some time and cost will be entailed. Otherwise, if incomplete information is used to make a decision quickly, we must put up with using a probability that will control a non-optimum system. We do not have complete accuracy for both the information that is available and decisions that are made. This is analogous to Heisenberg's uncertainty principle. This paper discusses the relation between information and decision in an optimal inventory process from this viewpoint.

Additionally, we introduce the general principle of balance. We thus possess two weapons namely the principle of optimality in dynamic programming and the principle of balance in a management information system. In the third section, this principle of a balance will be applied to the development of the relation between information and decision in optimal inventory processes. Then, problems regarding quantity approximation, time approximation, demand approximation, the criterion approximation and system structure approximation are summarized. The fourth section discusses the stability of the optimal inventory equation and presents a design for an optimal inventory system.

Finally, we point out that one source of imprecision stems from both randomness and fuzziness, and conclude with a discussion of some areas for further research.

Principle of Balance

The stochastic properties of quantum mechanics are based on the uncertainty principle. A balance relation is pointed out wherein it is theoretically impossible to measure with the same accuracy at the same time for a pair of quantities, called a conjugate quantity.

The phenomenon that is analogous to this principle in physics exists in many fields. Let us generally call this the principle of balance and discuss the relationship between this principle and several phenomena.

For example, the approximating of linear prediction theory due to Wiener leads to the problem of minimizing the quadratic form

$$E = \sum_{k=0}^{N} (a_k - \sum_{\ell=0}^{M} A_\ell a_{k-\ell})^2$$

over the real quantities A_ℓ , where the quantities a_k are given real numbers. E is the prediction error. The prediction error decrease and the structure complex increase when M is increased. It is an important practical question of decide how large to make M that balance the prediction error and the structure complex.

(1) Principle of Optimality in Dynamic Programming [1]

The principle of optimality in dynamic programming indicates that the optimal policy should harmonize the balance between costs involved in deciding present and future values on a new state reduced by its decision, because dynamic programming involves multi-stage decision processes. The information for the future is necessary in to make a decision in the present. The principle of optimality is an exact mathematical expression for this idea.

Let us assume Rth multi-stage decis'on processes. We shall be concerned with criteria possessing a structure which ermits us to focus our attention solely upon the past and present history of the process in a search for values of policies. Then, to construct the optimal policy of the Rth stage, whatever the initial state and initial decision are, the remaining (R - 1)th decisions must constitute an optimal policy with regard to the state resulting from the decision on the first stage. We must determine the first decision in order to determine the balance between gain in the first stage and gains in remaining (R - 1) stages.

(2) Principle of Balance in Information and Decision

If complete knowledge of the system is deemed necessary at any stage, then an appreciable time is usually required to accumulate this data. During this time, the system is oncontrolled. That is to say that time is one of the most valuable resources we have; it is unique in the fact that it cannot be reversed or replaced. It takes time to make decisions and then to implement those decisions. If, however, we make a decision quickly, using incomplete information about the system, there is a non-negligible probability that a non-optimal action will be taken. We cannot have complete accuracy in both information and control. This is the uncertainty principle in a management information system.

Information and Decision in Optimal Inventory Processes

This section discusses some applications of principle of balance in regard to information and decision in multi-stage stochastic inventory control processes. Multi-stage stochastic inventory control processes will be introduced in Section 4. At first, if we observe the exact inventory quantities, then we have the right decision and the optimum expected cost, but we must accordingly allow for the cost of more observation. This sort of approximation relates to the quantity aspects. Secondly, instead of keeping records and placing orders at each period, it may be better to observe and order at intervals of a few periods, even when this delay necessitates paying a penalty charge for getting items quickly. This type of approximation relates to the component of time. Also, there are some approximation problems in regard to determining demand information, optimum criterion and inventory system structure, etc.

(1) Approximation of Observation [7], [8]

A major problem in modern management is that of keeping records. However, sometimes, at a certain point, the cost of keeping records is greater than the gain that is obtained by using these records. These factors provide the motivation for a study of the approximation of observation of inventory quantity. It is necessary to decide on the degree of observation approximation that harmonizes with the observation cost and the gain obtained by using approximate information. We have obtained the following results, using both analytic and computational studies. [7, 8]

1) Optimal choices between degree of observation M and degree of policy N depend on the unit costs for this inventory process.
2) Inventory processes are as sensitive to M as to N .
3) Inventory processes are as sensitive to the backlogging problem as to the lost sales probler., etc.

We can determine the degree of approximation that balances the cost of observation and the total expected cost, if the approximate observation quantity is used.

(2) Approximation in Time [9], [10]

In introducing the basic optimal inventory equation, explicit use was made of the assumption that observations and orders are made at each period. However, this assamption may be questionable. Instead of keeping records in every period, it may be better to count the number of items when the supply is low, and even to pay a penalty charge for getting items quickly when the supply is very low. The problem that we want to study is that of determining the time to examine the number of items remaining in stock.

The results of analytic and computational studies are given in [10]. As we might have expected, the shortage and the total expected cost increase with increasing variability of time interval in decision. Thus, we can determine the time of observation and control that balances the cost of observation and the expected cost, which are obtained by using an approximate time.

(3) Approximation of Demand Information [11]

The first step away from completely deterministic demands and a step of considerable import, is the classical theory of probability with its introduction of random variables. We want to indicate the existence of high levels of uncertainty. We can consider the following three cases.

1) Stochastic problem; the case when the stochastic feature is known.
2) Adaptive problem; the case where the demand distribution contains unknown stochastic parameters.
3) Game theoretic problem; the case when the stochastic feature is unknown.

In [11], we have compared the solutions for cases when the probability density functions of demand are assumed to be exactly known, adaptively known and game theoretically known. The total expected cost increases as the completeness of information decreases.

(4) Approximation in Criterion [12]

The problem of establishing the inventory system effectiveness criterion is a very fundamental one.

Let us discuss the following problems of a multi-stage nature, namely average cost per period and probability criterion.

1) Multi-stage problem

We often say that we are planning for the next year, and that we wish to minimize the multi-stage expected costs or maximize the multi-stage expected profits. It is clear that managers need not plan for next year only and that, in fact, they must consider many years in advance. In this case, the optimal policy in one period does not always mean the optimal policy in multi-stage periods. However, under some assumption, the former coincides with the latter.

2) Average cost per period

In the stationary approach, we select a particular (s, S) policy, calculate the long-run costs based on this policy, and then select the policy variables so as to minimize long-run cost. Let the minimum cost be denoted by k. In the dynamic programming approach, the technique depends on the minimum cost function $C_n(x)$. If the interest rate is zero, then, as period n becomes infinite, $C_n(x)$ will tend toward infinity. It seems plausible that there will be some connection between $\lim_{n \to \infty} [C_n(x)]/n$ and k.

3) Probability criterion [12]

Let us discuss the criterion which minimizes the probability that the inventory over all stages exceeds a fixed level. The profits in the probability criterion are as follow. At first, this is simple, because we do not require an estimation of the cost functions. Secondly, we have the same policy characterized by the principle of constant stock level as the criterion of cost functions.

(5) System Structure Approximation [12]

Consider an inventory system that has many benefits. Under individual inventory control, each location puts in their orders separately and is concerned only with its own welfare. Under its centralized inventory control procedure, by contrast, quantity orders are made simultaneously for all locations in the network. There are immediate advantages and disadvantages to controlling such an inventory system centrally.

Since information about the entire supply network is recorded at a central location, decisions can be made effectively and expediently in emergencies, but the resulting decisions are more complex. An important question is the determination of how many benefits that are optimal in order to achieve centralized control.

Inventory System Design

Most of the authors who have written on the subject of inventory control have made the assumptions either that we have obtained or that we shall have information used to make the necessary decision. In this section, on the contrary, let us determine the kind and accuracy of the information, on the assumption that we know how to decide when to have some information. [3]

There are two types of costs with regard to information. One is the observation cost, which is entailed in obtaining information. The other is the error cost, owing to the approximation of information. Our aim is to minimize the sum of these costs. A model of our inventory control process is the multi-stage stochastic inventory problem. Let $L(y)$ be given by:

$$L(y) = \int_0^y h(y-\xi)\phi(\xi)d\xi + \int_y^\infty p(\xi-y)\phi(\xi)d\xi \qquad (1)$$

Where $L(y)$ represents the sum of the expected inventory cost $\int_0^y h(y-\xi)\phi(\xi)d\xi$ and the expected penalty cost $\int_y^\infty p(\xi-y)\phi(\xi)d\xi$ in each period, and $\phi(\xi)$ is the demand probability density in each period, given that, at the beginning of a period, the sum of the initial inventory on hand and the stock to be received in a period is y. We define the functional $f_n(x)$ as the total expected discounted cost over n periods, where x is the inventory on hand at the beginning of the first period. We have observed that $f_n(x)$ can be written for all x in the following functional equation.

$$f_n(x) = \min_{y \geq x} [c(y-x)+L(y)+\alpha\int_0^\infty f_{n-1}(x-\xi)\phi(\xi)d\xi] \qquad (2)$$

In (2) $c(y)$ is the ordering cost and α denotes the discount factor $(0 < \alpha < 1)$. At first, we shall review the stability of the inventory process which is fundamental to design issues. In the following, we introduce the inventory processes design and the inventory policies control problems.

DISCUSSION

We have shown that the environment for approximation of observations and policies affects our inventory processes.

Much of the decision making in the real world takes place in an environment in which the goals, the constraints and the consequences of possible actions are not known precisely. To deal quantitatively with imprecision, the traditional approach is to employ the concepts and techniques of probability theory and, more particularly, the tools provided by decision theory, control theory and information theory. This is now questionable especially in view of the developments in the field of fuzzy sets theory [4], [5].

There is a differentiation between randomness and fuzziness, with the latter being a major source of imprecision in many decision processes. By fuzziness, we mean a type of impression which is associated with fuzzy sets, that is, classes in which there is no sharp transition from membership to nonmembership. There are many facets of the theory of decision making in a fuzzy environment which require more thorough investigation. This is also the case with inventory systems.

REFERENCES

1. R. Bellman, Dynamic Programming (Princeton University Press 1957).
2. B.E. Bellman, Introduction to the Mathematical Theory of Control Processes (Academic Press, New York 1971).
3. R.L. Ackoff, and M.W. Sasieni, Fundamentals of Operations Research, Chap. 14. (John Wiley & Sons, Inc. 1968).
4. R.E. Bellman and L.A. Zadeh, Decision-making in a Fuzzy Environment, Management

Science, 17, 4 (1970).

5. J. Kacprzyk, and P. Staniewski, Long-term Inventory Policy-making through Fuzzy Decision-making Models, Fuzzy Sets and Systems, 8, 2 (1982).
6. T. Odanaka, A Study of Multi-stage Inventory Control, Ph.D. Thesis of T.I.T. (1968).
7. T. Odanaka, Analytic and Computational Studies of Optimal Inventory Processes I, USC Technical Report, USCEE-352 (1969).
8. T. Odanaka, On Approximation on Quantity in Optimal Inventory Processes, Policy and Information, 5, 1 (1981).
9. T. Odanaka, Analytic and Computational Studies of Optimal Inventory Processes II, USC Technical Report, USCEE-353 (1969).
10. T. Odanaka, On Approximation in Time in Optimal Inventory Processes, Journal of Operations Research Society of Japan, 18, 12 (1975).
11. T. Odanaka, and S. Maruyama, Analytical and Computational Solution of Adaptive Inventory Processes, Journal of Mathematical Analysis and Application (1985).
12. T. Odanaka, Optimal Inventory Processes, Katakura Lib. (1985).

CLASSIFICATION METHOD USING MULTICRITERION OPTIMISATION : APPLICATION TO THE STOCK EXCHANGE.

by **LITWAK R.G ; LAURENT R ; POVY L ; HUREL D ;**
Centre d'Automatique de Lille
USTL.Flandre.Artois
59655 Villeneuve d'ascq cedex -FRANCE

ABSTRACT :

This paper propound a preference model and a decision model for an economic system. With the multicriterion optimisation, we develop a method permitting the choice of stock portfolio in stock exchange. In the first part, we compare two fuzzy numbers, representative of the value of two actions according to a criterion, and we introduce the threshold notion. In the second part, we compare two classing result from two criterions ; a new classing will be deduct of this comparison. This one give the composition of a portfolio.

I INTRODUCTION

The stock exchange is too complex economic system to build an elaborate model. Nevertheless, with a simple model, we can have a better understanding of the system ; and so deduct a best adapted command.

There is two types of models :

- Economic model : with several parameters as social, political, financial events.
- Technical model : based on the quotations, these models give the value of the Beta coefficient, the index ....

The first one, too much subjective, must be used with a lot of precautions. The second one contains a part of the information wich circulates in the market. The estimations given by the technical models are often biaised ; since they are obtained by the study of the past. It is important to utilize a method which integrate this error.

These different models give a set of criterions which allow to valuate the stock exchange.

Let $C=\{C_1,C_2,...,C_m\}$ the criterion set.

$A=\{A_1,A_2,...,A_n\}$ the possible set of stocks on the system

To integrate the error of estimation, we consider an uncertainty domain containing the value. So all the values become fuzzy numbers

II COMPARISON OF TWO FUZZY NUMBER

In this part, we develop a preference model allowing the comparison of two fuzzy numbers. In our application, each criterion is applied on each stock. So a criterion matrix $C_i(a_j)$ C_{ij} can be created. With the notion of fuzzy numbers, the criterions become :

$[C_i^-(a_j),C_i^+(a_j)]$; where $[C^-,C^+]$ represent the limits of a uncertainty interval. The domain of fuzzy number, representative of the criterion, integrate the error of estimation. We must add to this error, the possibility of a future event unforeseeable by the past. For that reason, we introduce a threshold notion : S_i. A stock (a_j) will be better than an other stock (a_k), for a criterion (C_i), if the value of the criterion applied to the stock (a_j) is better than the value obtained with the second stock (a_k) ; and if the difference is higher than the threshold S_i.

If $C_i(a_j) > C_i(a_k) + S_i$ => a_j is prefered to a_k for C_i

In this case, a perturbation P, due to an unforeseeable event, will not change the choice, if his amplitude is not higher than the threshold S_i.

if $P < S_i$ we have :
$C_i(a_j) > C_i(a_k) + P$ => a_j is always prefered to a_k
for the criterion C_i.

We see immediately than if :
$C_i(a_j) < C_i(a_k) + S_i$
$C_i(a_k) < C_i(a_j) + S_i$.
Then, a_j is not prefered to a_k and a_k is not prefered to a_j. To describe this possibility, we introduce the indifference's relation ; which is noted : I .
In order, to compare two stocks a_j and a_k, by using the criterion C_i, we have consequently three choice relations :

- Preference (P)
- Indifference (I)
- No preference (NP)

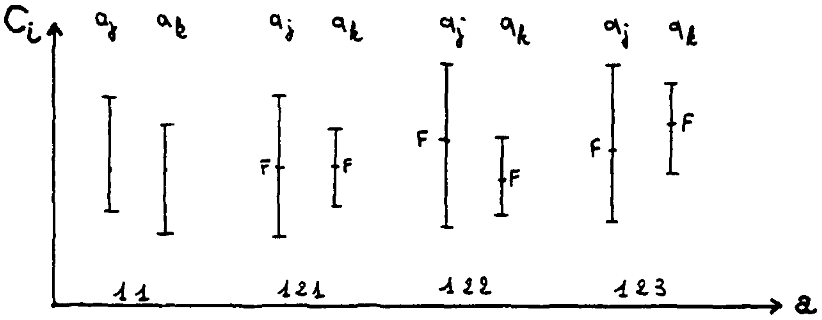
We define the different relations as following :

1 $C_i+(a_j)-C_i+(a_k) >$ S_i
11 $C_i-(aj)-C_i-(a_k) >$ S_i => a_j P a_k
12 $C_i-(a_j)-C_i-(ak) <=$ Si

Considering :
$M = ((C_i+(a_j) + C_i-(a_j))-(C_i+(a_k) + C_i-(a_k)))/2$

We can use a function of M, noted : F. This function (f) depends on several parameters (uncertainty domain, threshold, ...).

121	$-S_i <=$	$F <= S_i$	=>	a_j I a_k
122		$F > S_i$	=>	a_j P a_k
123		$F < -S_i$	=>	a_j NP a_k



Representation of the different cases

2 $C_i+(a_j)-C_i+(a_k) <$ $-S_i$

The obtained results are the same than the precedent, if we change j and k.

3	$-S_i <= C_j+(a_j)-C_j+(a_k) <=$	S_i			
31	$-S_i <= F <=$	S_i	=>	a_j I	a_k
	$F >$	S_i	=>	a_j P	a_k
	$F <$	$-S_i$	=>	a_j NP	a_k

The method consist in comparing the n actions two by two, for the criterion C_i. A preference matrix $Pc_i(j,k)$ (n*n dimension) can be built. To use this matrix, we must code the preference relations.

We take the following coding :

if a_j	P	a_k	=>	$Pc_i(j,k) = 1$
if a_j	I	a_k	=>	$Pc_i(j,k) = 0$
if a_j	NP	a_k	=>	$Pc_i(j,k) = -1$

Every criterion will have a matrix n*n which display the comparison between the stocks.

For one criterion, two stocks are compared owing to the rows of the matrix Pc_i. The row j shows the preference for the stock j with regard to the others. The value of the preference for this stock is given by the sum of the row.

Let us build the vector :

$V{:}om_i(k) = Pc_i(j,k)$

$Vsom_i(k)$ represent a performance measure of the stock a_k. We would terminate this study, if the found classification was the same for all criterion C_i : this is improbable. So we introduce a method which compare two contradictory classing. This method will give the value of all the stocks for the criterions set and then a classing.

III COMPARISON BETWEEN TWO CRITERIONS

In this part, we develop a decision model allowing the comparison of two criterions. This method consist in setting the criterions in a hierarchical order, beginning by the most important for the investor, and ending by the less one. With two classing, related to two criterions, we build a new classing. We take the two most important criterions, then the second and third, and so on. To be able to define a global classing with two contradictory primary classing, we introduce the criterion preference coefficient, written CL_i.

Let a_j and a_k be two stocks, C_i and C_{i+1} be two criterions with C_i prefered to C_{i+1}. The stocks, a_j and a_k have each one a value for each criterion : $Vsom_i(a_j)$,$Vsom_{i+1}(a_j)$ for the stock a_j ; and $Vsom_i(a_k)$,$Vsom_{i+1}(a_k)$ for the stock a_k. More exactly, we normelize the vectors $Vsom_i(a_j)$ by doing the following tranformation :

$$Vsom_i(a_j) = (Vsom_i(a_j)-Vsom_i\ min)/(Vsom_i\ max-Vsom_i\ min)$$

With $Vsom_i$ min and $Vsom_i$ max, the minimum and the maximum value of the vector $Vsom_i(a_j)$, for the criterion C_i.

In order, to establish the new classing, we make the difference between the two values of the two stocks for a criterion :

$$D_i = Vsom_i(a_j)-Vsom_i(a_k)$$

if $D_i > 0$ => aj P ak

if $D_{i+1} < 0$ => ak P aj

This two criterions do not permit us to choose between the stocks a_j and a_k. A solution consist in saying that the second criterion is less important than the first one. Then we multiply the second (D_{i+1}) by the criterion preference coefficient (CP_i) ; with CP_i less than CP_{i+1}. The new preference relation becomes :

$$D_{i+1} \Rightarrow D_{i+1} * CP_i$$

So we have the following relations :

1) Obvious relations :

$D_i > 0$ and $D_{i+1} >= 0$	=>	a_j P a_k
$D_i < 0$ and $D_{i+1} <= 0$	=>	a_j NP a_k
$D_i = 0$ and $D_{i+1} = 0$	=>	a_j I a_k
$D_i = 0$ and $D_{i+1} > 0$	=>	a_j P a_k
$D_i = 0$ and $D_{i+1} < 0$	=>	a_j NP a_k

2) Solutions depending of CP_i :

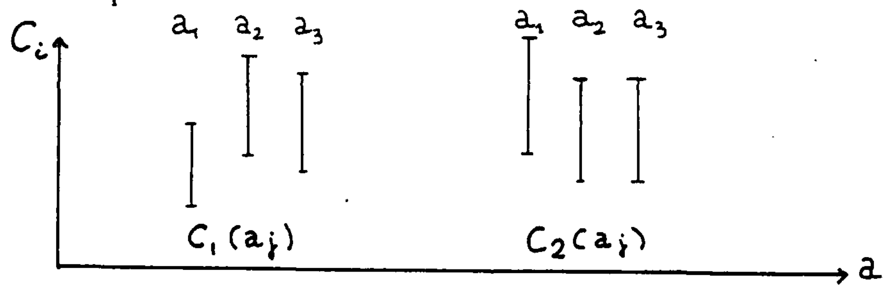
$D_i > 0$ and $D_{i+1} < 0$ or $D_i < 0$ and $D_{i+1} > 0$

$D_i + D_{i+1}*CP_i > 0$ => a_j P a_k
$D_i + D_{i+1}*CP_i < 0$ => a_j NP a_k
$D_i + D_{i+1}*CP_i = 0$ => a_j I a_k

By comparing n stocks with themselves, we obtain a preference matrix P_i (n*n). And then, we calculate the sum vector $VS_i(k)$; it gives so a new classing which use the two criterions C_i and C_{i+1}. With the criterion C_{i+1} and C_{i+2}, we obtain the sum vector $VS_{i+1}(k)$. The result of VS_i and VS_{i+1} will be compared with C_{i+3} and will give $VS_{i+2}(k)$; and so on, until $VS_m(k)$. At least,the sum vector will give us the definitive classing for the criterions set.

Exemple :

let $A=\{A1,A2,A3\}$
$C=\{C1,C2\}$
$S_i=0$



$$Pc_1 = \begin{pmatrix} 0 & -1 & -1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \Rightarrow Vsom_1 = \begin{pmatrix} -2 \\ 2 \\ 0 \end{pmatrix}$$

$$Pc_2 = \begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \Rightarrow Vsom_2 = \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix}$$

let $CL_i=0$

$$P_1 = \begin{pmatrix} 0 & -1 & 1 \\ 1 & 0 & 1 \\ -1 & -1 & 0 \end{pmatrix} \Rightarrow VS_1 = \begin{pmatrix} 0 \\ 2 \\ -2 \end{pmatrix}$$

IV CONCLUSION

The use of this method, on the stocks of the stock exchange of Paris, has a real increase of the portfolio profit with regard to the method of simple investing. Our work, at the present time, is oriented to the amelioration of these models ; with the function F, as also on the determination of the different thresholds.

BIBLIOGRAPHY:

[1] BENAYOUN R ; De MONTGOLFIER J ; TERGUY J
Linear programming with multiple objective functions : STEM
MATHEMATICAL PROGRAMMING dec 1971 n°12 page 366-375

[2] FOURGEAUD ; LENCLUD ; SENTIS
Critere de choix en avenir incertain
R.I.R.O 1968 n°14 page 09-20

[3] KEENEY R
Utility fonctions for multiattributed consequences
MANAGEMENT SCIENCE Jan 1972 n°5 page 276-287

[4] KOLKI J
Defectiveness of weighting method in multicriterion optimisation of stuctures
APPLIED NUMERICAL METHODS 1985 n°1 page 333-337

[5] ROY B
ELECTRE III : un algorithme de classements fonde sur une representation floue des preferences en presence de critere multiple
CAHIER DU CENTRE D'ETUDES ET DE RECHERCHE OPERATIONNELLE
1978 n°1 page 03-24

[6] STADER W
A survey of multicriteria optimizatioin or vector maximum problem
JOURNAL OF OPTIMIZATION 1979 n°1 page 1-51

SYSTEMES STOCHASTIQUES ET QUANTIQUES

STOCHASTIC AND QUANTUM SYSTEMS

Logarithmic Transformations with Applications in Probability and Stochastic Control

Wendell H. Fleming

Division of Applied Mathematics, Brown University, Providence,
Rhode Island 02912.

Abstract

We are concerned with a class of problems described in a somewhat imprecise way as follows. Consider a linear operator of the form L + V(x), where L is the generator of some Markov process x_t and the "potential" V(x) is some real-valued function on the state space of x_t. We are interested in probabilistic representations for solutions u(t,x) of the evolution equation

$$(1) \qquad \frac{\partial u}{\partial t} = Lu + V(x)u, \ t > 0$$

with initial data at t = 0. The Feynman-Kac formula gives a well-known stochastic representation for u(t,x). We seek a different probabilistic representation for $I = -\log u$, if u(t,x) is a positive solution to (1). In this representation the operator L is replaced by another generator $\tilde{L}_t$ (perhaps time dependent), chosen to solve a certain stochastic control problem. The dynamic programming equation for this stochastic control problem is

$$(2) \qquad \frac{\partial I}{\partial t} = H(I) - V(x), \text{ where}$$
$$H(I) = -e^{I}L(e^{-I}).$$

Another way to view the change of generator from L to $\tilde{L}_t$ is by change of probability measure through conditioning.

Next suppose that the state space of x_t is euclidean R^n, that

$$L = L_\epsilon, \ u = u^\epsilon, \ I^\epsilon = -\epsilon \log u^\epsilon \text{ and}$$
$$H_\epsilon(I) = -e^{I}L_\epsilon(e^{-I}).$$

Under various assumptions it turns out that $I^\epsilon \to I^0$ as $\epsilon \to 0$,

$$\lim_{\epsilon \to 0} \epsilon H_\epsilon(\epsilon^{-1} I) = H_0(x, I_x)$$

where I_x is the gradient, and that I(t,x) is a viscosity solution of the first-order partial differential equation

$$\frac{\partial I}{\partial t} = H_0(x, I_x).$$

When x_t is a nondegenerate diffusion on R^n, then L is a second order elliptic partial differential operator. In this case, the logarithmic transmation provides an analytical approach to large deviations questions of Ventsel-Freidlin type, and for more precise results in the form of asymptotic series expansions of I^ϵ in powers of ϵ. The logarithmic transformation technique is also of use to study certain asymptotic problems in which $u^\epsilon(t,x)$ obeys a nonlinear parabolic partial differential equation.

References

1. G. Barles and B. Perthame, Exit time problems in optimal control and vanishing viscosity, to appear in SIAM J. Control Optimiz.

2. P. DuPuis, Minimizing exit probabilities: a large deviations approximation, Brown Univ. LCDS Report, (1987).

3. L.C. Evans and H. Ishii, A PDE approach to some asymptotic problems concerning random differential equations with small noise intensities, Ann. Inst. H. Poincare Analyse NonLineaire 2 (1985) 1-20.

4. L.C. Evans and P.E. Souganidis, A PDE approach to geometric optics for certain semilinear parabolic equations, preprint.

5. W.H. Fleming, Exit probabilities and optimal stochastic control, Appl. Math. Optimiz. 4 (1978) 329-346.

6. W.H. Fleming, Stochastic calculus of variations and mechanics, J. Optimiz. Th. Appl. 41 (1983) 55-74.

7. W.H. Fleming, Logarithmic transformations and stochastic control, in Springer Lecture Notes in Control and Info. Sci. No. 42 (eds. W.H. Fleming and L.G. Gorostiza) Springer-Verlag (1982) 131-141.

8. W.H. Fleming, A stochastic control approach to some large deviations problems, in Springer Lecture Notes in Math no. 1119 (eds. I. Capuzzo Dolcetta, W.H. Fleming and T. Zolezzi), (1984) 52-66.

9. W.H. Fleming, Controlled Markov processes and viscosity solutions of nonlinear evolution equations, Lezione Fermiane (1986), Accademia Nazionale dei Lincei, Scuola Normale Superiore, Pisa.

10. W.H. Fleming and S.K. Mitter, Optimal control and nonlinear filtering for nondegenerate diffusion processes, Stochastics 8 (1982) 63-77.

11. W.H. Fleming and H.M. Soner, Asymptotic expansions for Markov processes with Levy generators, to appear in Applied Math. and Optimiz.

12. W H. Fleming and P.E. Souganidis, PDE-Viscosity solution approach to some problems of large deviations, Annali Scuola Normale Sup. Pisa Classe Sci., Ser IV 23 (1986) 171-192.

13. W.H. Fleming and P.E. Souganidis, Asymptotic series and the method of vanishing viscosity, Indiana Univ. Math. J. 35 (1986) 425-447.

14. C.J. Holland, A new energy characterization of the smallest eigenvalue of the Schrödinger equation, Comm. Pure Appl. Math. 30 (1977).

15. S.-J. Sheu, Optimal control and its application to large deviation theory, Brown Univ. Ph.D. Thesis 1983.

16. S.-J. Sheu, Stochastic control and principal eigenvalue, Stochastics 11, (1984) 191-211.

MACROSCOPIC PROPERTIES
OF
DISCRETE DIFFUSIONS

by

E. Benoit*, B. Candelpergher**, C. Lobry**

The mathematical Wiener process, and more generally mathematical diffusion processes, are mathematical idealizations of physical processes like Brownian motion for instance . The starting point in the definition of the Wiener process is the definition of the random walk :

$$\text{R.W.}(dt) \quad \begin{cases} \xi_{t+dt} = \xi_t + Z_t\sqrt{dt} & t \in \{\, 0, dt, 2dt, \dots, pdt, \dots ndt = 1 \} \\ \xi_0 = 0 \end{cases}$$

where Z_t is a sequence of independent random variables taking values ±1 with equal probabilities . This is certainly a simple mathematical object but not a good idealization because it contains a parameter dt and , *à priori* , there is no universal way to fix it .

As it is well known, the usual way to deal with this difficulty is to consider the whole familly R.W.(dt) (dt > 0) and, because we want to idealize physical random motions in which the elementary step is very small compare to the scale where the phenomenum is observed, we take the limit when dt → 0 . This limit is a mathematical object known as the **Wiener process**. Mathematically speaking the Wiener process is a probability measure on the infinite dimensional vector space : $\mathbb{R}^{[0\ ,\ 1]}$. Because of the great cardinality of this space a probility measure is no longer a simple object and its definition requires the knowledge of most of the technicalities of measure theory . The only way to avoid all those technicalities is to work only on probability spaces of finite cardinality . This is the case of the random walk R.W.(dt) which is defined on the soace $\{-1\ .\ +1\}^n$. But we lose universalitu of the wiener process . How can we recover it ?

Define the **Wiener Walk** (W.W) as R.W.(dt) for dt **infinitesimal** and look for its **macroscopic** properties .

The real number dt considered here is infinitesimal in the **formal sense** of Non Standard Analysis (N.S.A.) . A macroscopic observation of the process is an observation which is not able to distinguish differences at the microscopic level : For instance if the position of the process at time t is x we consider that our mesurement is not able to give the exact value of x but merely any value which is infinitely close to x ; any such a value is idealized by the **shadow** of x, which is the **standard** real number infinitesimally close to x . Thus the property :

"The shadow of ξ_t is positive"

is a sentence which makes sense in the language of N.S.A. (but not in the conventional language ant thus is called **external**) and expresses a property of W.W. which makes no reference to dt . It does not depend on dt provided it is infinitesimal . By this way we recover universality of the idealization .

This mathematical model in which the law of the process is defined at the microscopic level (dt) and is observed at a macroscopic one fits very well with the Brownian motion in which we actually observe at a very large scale (say 10^{-6}m for position and 10^{-2} sec for time) the consequences of about $10^{21}-10^{22}$ kicks per second by the molecules of a gaz on the Brownian particle .

In this approach all the technicalities associated to measure theory are suppressed and replaced by those which are associated to the use of the richer language of N.S.A. Where is the benefit ? The benefit is in the fact that we need very few elements of N.S.A. , without comparison with what we need from mesure theory . This was recognised by E. NELSON [1] . In this book, among other things, he establishes in less than 80 pages (including the necessary rudiments of N.S.A.) the external equivalents of all the essential properties and caracterisations of the Wiener process . In our paper [2] we have extended the approach of NELSON to the more general processes :

$$\xi_{t+dt} = \xi_t\, b(\xi_t,t)dt + s(\xi_t)\, Z_t\sqrt{dt} \qquad t \in \{\, 0, dt, 2dt,\dots,pdt,\dots ndt = 1\}$$
$$\xi_0 = 0$$

In this lecture we shall explain :

1) What means "almost sure" on a finite set with uniform probability .

2) What means "(nearly)continuous" for a discrete mapping from $\{\, 0, dt, 2dt,\dots,pdt,\dots ndt = 1\}$ to $\mathbb{R}$ and why it is a macroscopic concept .

3) How one can prove the following :

Theorem : Consider the process defined by :

$$\xi_{t+dt} = \xi_t\, b(\xi_t,t)dt + s(\xi_t)\, Z_t\sqrt{dt} \qquad t \in \{\, 0, dt, 2dt,\dots,pdt,\dots ndt = 1\}$$
$$\xi_0 = 0$$

where dt is an infinitesimal, Z_t is a sequence of independant random variables taking values in ±1 with equal probabilities and the mappings b(.,.) and s(.) are standard continuously differentiable, bounded with their derivative . Then "almost every" trajectory is "(nearly) continuous" .

REFERENCES

[1] **E. NELSON** : RADICALLY ELEMENTARY PROBABILITY THEORY, Annals of Math.Studies. 117 Princeton University Press 1987.

[2] **E. BENOIT , B. CANDELPERGHER, C. LOBRY** : DIFFUSIONS DISCRETES et MECANIQUESTOCHASTIQUE, *Centre de Mathematiques Appliquées, Ecole des mines Sophia Antipolis et Département de* Mathématiques Université de Nice , prépublication 160, Octobre 1987 .

(*) E. Benoit : Centre de Mathématiques appliquées, Ecole des Mines, Sophia Antipolis, 06565 VALBONNE Cedex France.

(**)B. Candelpergher, C. Lobry : Département de Mathématiques, Université de Nice, Parc Valrose, 06084 NICE Cedex France.

LOCAL CONTROLLABILITY OF GENERALIZED QUANTUM MECHANICAL SYSTEMS

T.J. Tarn[+], John W. Clark[++] and Garng M. Huang[+++]

[+] Department of Systems Science and Mathematics, and Center of Robotics and Automation, Washington University, St. Louis, Missouri 63130, U.S.A.

[++] Department of Physics, and McDonnell Center for the Space Sciences, Washington University, St. Louis, Missouri 63130, U.S.A.

[+++] Department of Electrical Engineering, Texas A & M University, College Station, Texas 77843, U.S.A.

ABSTRACT

The concept of local controllability is investigated for non-relativistic quantum systems. Sufficient conditions will be sought such that the solution of the controlled Schrodinger equation can be guided, over a short time interval, to any chosen point in a suitably prescribed neighborhood of the solution in the absence of control. Evolution equations which are linear in the controls but nonlinear in the quantum state ψ are considered. Our formulation and analysis will (for the most part) run parallel to those of Hermes.

I. INTRODUCTION

In recent years, there has been a growing interest in the system theoretic problems of filtering and control of quantum mechanical systems. Several noteworthy efforts exists: (i) Tarn, Huang and Clark [1] and van der Schaft [2] have explored the formal basis for the modelling of quantum mechanical control systems. (ii) Clark, Tarn and their associates [3-6] have obtained results on quantum nondemolition filtering problem. (iii) Belavkin [7] has investigated the measurement and control problem in quantum dynamical systems. (iv) Pierce, Dahleh and Rabitz [8] have studied the optimal control problem of quantum mechanical systems. (v) Butkovskiy and collaborators have discussed the control of quantum objects in broad terms and have set forth general conditions for controllability of pure quantum states [9-11].

To the authors' knowledge very little has been published in the way of mathematically definitive results on the controllability of quantum systems. In [12] the authors are able to establish a series of global controllability conditions for the Schrodinger equation which is linear in state and linear in the external controls by extending the geometric approach as implemented by Sussmann and Jurdjevic [13,14], Krener [15], Brockett [16], Kunita [17] and others.

In the present contribution, we shall consider evolution equations which are linear in controls but nonlinear in the quantum state; in this case the work of Hermes [18] is extended to obtain conditions for local controllability along an unguided reference solution.

II. PROBLEM FORMULATION WITH NONLINEAR GENERATORS

In adapting Hermes' work [18] to our ends, it is convenient to think in terms of the x representation [19]. Thus the state vector $\xi \in H$ will be represented by the wave function $\xi(x) \in L_2(R^n)$, where $x \in R^n$ stands (ordinarily) for the set of spatial coordinate variables associated with the quantum system. (More generally, x may stand for any complete set of compatible variables [19] built from the position and momentum variables. Spin and other internal degrees of freedom can be incorporated by essentially trivial modifications.) Now, let us define a class of operators H in H which are supposed to be skew-Hermitian (norm preserving) and time independent and to have, in the x representation, the mode of action

$$(H\xi)(x) \equiv H\xi|_x = \sum_{\lambda=1}^{p} f_{\lambda,1}((H_{\lambda,1}\xi)(x))\dots f_{\lambda,q}((H_{\lambda,q}\xi)(x)). \qquad (1)$$

Here, p, q are some integers, the $H_{\lambda,\mu}$ ($\lambda = 1,\dots,p$; $\mu = 1,\dots,q$) are closed, skew-Hermitian linear operators acting in H, and the mappings $f_{\lambda,\mu}$: $C^1 \to C^1$ are real analytic. (By the last requirement we mean that $f_{\lambda,\mu}(w)$ is a real analytic function of its argument w, this argument in itself being generally complex, $w \in C^1$. Also, in expression (1), $f_{\lambda,\mu}(w)f_{\lambda',\mu'}(w')$ is to be interpreted as the usual product of complex functions.) Throughout the current section, the generators $H_0,\dots,H_r$ entering the "controlled Schrodinger equation" will be assumed to be of this more general form. Thus, while $H_0,\dots,H_r$ are still taken skew-Hermitian, they need not be linear--although the linear case is certainly included.

We shall further assume that a unique local solution exists for the initial value problem

$$\frac{d}{dt}\psi_t = \left[H_0 + \sum_{\ell=1}^{r} u_\ell(t)H_\ell\right]\psi_t, \qquad \psi_{t=0} = \phi \in H , \qquad (2)$$

posed by the Schrodinger equation so generalized, the admissible controls u_ℓ now being real, analytic, bounded functions of t. To establish that this is a viable assumption, we note that it is automatically fulfilled within the framework of [12], provided ϕ belongs to the analytic domain D_ω; moreover, in Ref. 20 it has been shown to be valid for a certain relevant class of partial differential equations. On the other hand the formulation of general conditions on $H_0 + \sum u_\ell H_\ell$ for the existence of a unique local solution of (2) awaits further mathematical developments.

Our next task is to specify the Lie bracket appropriate to the (generally) infinite-dimensional, (generally) nonlinear control problem (2), wherein the H_k, $k=0,\dots,r$ are of type (1). First, we appeal to the chain rule to define a sort of derivative operator, DH, corresponding to an operator H of that type:

$$((DH\xi)(x))\zeta(x) = \sum_{\lambda=1}^{p}\sum_{\mu=1}^{q} f_{\lambda,1}((H_{\lambda,1}\xi)(x))\dots f_{\lambda,\mu-1}((H_{\lambda,\mu-1}\xi)(x))$$

$$\cdot f'_{\lambda,\mu}((H_{\lambda,\mu}\xi)(x))f_{\lambda,\mu+1}((H_{\lambda,\mu+1}\xi)(x))f_{\lambda,q}((H_{\lambda,q}\xi)(x))(H_{\lambda,\mu}\zeta)(x), \tag{3}$$

where $\zeta \in \mathcal{H}$ and $f'(w)$ is the derivative of $f(w)$ with respect to its argument. The Lie bracket of two operators H, K of the indicated class is then specified by

$$([H,K]\xi)(x) \equiv [H,K]\xi|_x = ((DH\xi)(x))(K\xi)(x) - ((DK\xi)(x))(H\xi)(x), \tag{4}$$

to apply $\forall\ \xi \in \mathcal{H}$ and $\forall\ x$. Again we shall employ the notation $\mathrm{ad}_H K = [H,K]$, $\mathrm{ad}_H^{\nu+1} K = [H,\mathrm{ad}^{\nu} K]$, $\nu = 1, 2,\dots$; also, $\mathrm{ad}_H^0 K \equiv K$. The prescription (4) for the Lie product is obviously consistent with that of [12], for, if H and K are linear, $[H,K] = HK - KH$ as in [12].

Remark 1. The above definitions and specifications are tenable even if H and $H_{\lambda,\mu}$ of (1) are not skew-Hermitian (or even if skew-Hermiticity is not a meaningful concept). As is well known, skew-Hermiticity of the generators of time displacement is an indispensible requirement in conventional quantum theory, where it is necessary for the probability interpretation of ψ_t. On the other hand, there are circumstances in which one may be led to drop this requirement, namely, (i) in approximate treatments of the Schrodinger equation designed to yield simple pictures of complicated phenomena involving many degrees of freedom, and (ii) in radical revisions of conventional quantum theory aimed at a more fundamental description of the microscopic world. The optical model of nuclear reactions, [21] wherein a complex potential is introduced to simulate the effects of inelastic processes, is a good example of circumstance (i), while the hadronic theory proposed by Santilli [22] suffices to illustrate possibility (ii). Obviously, in the latter context new interpretations as well as a new formal apparatus (see, e.g., Ref. 23) must accompany the enlarged mathematical framework.

Remark 2. The message of this comment is similar to that of Remark 1, except that the subject is nonlinearity of the generators $H_0, \dots, H_r$ rather than violation of their skew-Hermiticity. Conventional quantum mechanics is necessarily a linear theory, in that the superposition principle is an essential property. Specifically, linearity of the operators $H_0, \dots, H_r$ is required to maintain this property. But again one might agree, either (i) in the framework of approximation methods, or (ii) in fundamental extensions of quantum theory, to sacrifice linearity. The Hartree-Fock approximation [19,21] of atomic and nuclear physics furnishes a prominent example of a nonlinear approximation to the conventional quantum description. On the other side of the coin, nonlinear quantum theories at the first-principles level have been considered by a number of authors; for example, Wigner [24] has suggested that a resolution of the mysteries associated with "collapse of the wave packet" might be sought in terms of such a theory. [25]

III. GENERALIZED DECOMPOSITION THEOREM

Consider the system (2), wherein it is assumed that $\phi \in D \equiv \bigcap_{k=0}^{r} \mathrm{dom}\, H_k \neq$ null set. Let $V_t(\phi) \in D$ denote the solution (evaluated at time t) of the associated *reference problem*

$$\frac{d}{dt}\eta_t = H_0\eta_t \quad , \quad \eta_0 = \phi. \tag{5}$$

This problem corresponds to free evolution of the quantum system, the external controls being turned off; accordingly $\eta_t = V_t(\phi)$ will be referred to as the *homogeneous reference solution*. Treating ϕ, rewritten ζ, as an arbitrary element of the allowed domain D, we obtain a mapping $\zeta \to V_t(\zeta)$, which in general defines a nonlinear operator. (We note that in the special case that the generator H_0 is linear, $V_t(\zeta)$, which traces an integral curve of the vector field H_0, serves to define a *linear* evolution operator V_t. However, in the nonlinear setting of the present analysis, we are strictly not allowed to divorce operator from operand, since an operator of class (1) generally depends on the point of H at which it acts.) The differential of the mapping $\zeta \to v_t(\zeta)$, to be denoted $DV_t(\zeta)$, is also (generally) a nonlinear operator. One may loosely interpret $DV_t(\zeta)$, as the derivative of the object $V_t(\zeta)$, a state vector, with respect to its argument, which is again a state vector. By $DV_t(\zeta)|_x$ we will mean the differential of the (wave function) → (wave function) map $\zeta(x) \equiv \zeta|_x \to V_t(\zeta)|_x$.

Definition 1. A complex-valued function g: $t \to g(t) = g_1(t) + ig_2(t)$ is said to be *complex analytic* in the variable t, where $t \in R^1$, if the functions g_1 and g_2 are real analytic in t.

Theorem 1. (Generalized Decomposition Theorem (cf. Refs. 18,26)). Let ζ be an arbitrary element of the common domain D of the operators $H_0, \ldots, H_r$, and suppose that (i) the maps $t \to V_t(\zeta)|_x$ and $t \to DV_t(\zeta)|_x$ are complex analytic in t for all x and (ii) the differential $DV_t(\zeta)$ converges in the strong operator topology to the identity operator id, as $t \to 0^+$. Then, a sufficient condition for

$V_t(W_t(\phi))$ to provide a solution of the <u>controlled</u> dynamical problem (2), is that $W_t(\phi)$ satisfy

$$\frac{d}{dt}\xi_t = \sum_{\nu=0}^{\infty}\frac{(-t)^{\nu}}{\nu!}\left[ad^{\nu}_{H_0}\sum_{\ell=1}^{r}u_{\ell}H_{\ell}\right]\xi_t,\ \xi_0 = \phi \in D. \tag{6}$$

If $DV_t(\zeta)$ is one-to-one, the state condition is also necessary.

<u>Proof</u>. A necessary and sufficient condition for $V_t(W_t(\phi))$ to be a solution of (2), given that $V_0(W_0(\phi)) = W_0(\phi) = \phi$, is $H_0V_t(W_t(\phi)) + \sum_{\ell=1}^{r}u_{\ell}H_{\ell}V_t(W_t(\phi))$

$$= \frac{d}{dt}V_t(W_t(\phi)) = \frac{\partial}{\partial t}V_t(\zeta)|_{\zeta=W_t(\phi)} + DV_t(\phi)\frac{d}{dt}W_t(\phi) \quad . \tag{7}$$

Since by definition $V_t(\zeta)$ must satisfy the differential equation $\partial V_t(\zeta)/\partial t = H_0V_t(\zeta)$, where ζ may be regarded as an <u>independent variable</u> so far as the time derivative is concerned, the initial terms in the first and last members of (7) cancel. Thus condition (7) may be distilled to

$$DV_t(\phi)\frac{d}{dt}W_t(\phi) = \left[\sum_{\ell=1}^{r}u_{\ell}(t)H_{\ell}\right]V_t(W_t(\phi)) \quad . \tag{8}$$

The crucial step is to prove that, $\forall\ \zeta$ and $\forall\ x$,

$$DV_t(\zeta)\sum_{\nu=0}^{\infty}\frac{(-t)^{\nu}}{\nu!}\left[ad^{\nu}_{H_0}\sum_{\ell=1}^{r}u_{\ell}(t)H_{\ell}\right]\zeta|_x = \left[\sum_{\ell=1}^{r}u_{\ell}(t)H_{\ell}\right]V_t(\zeta)|_x. \tag{9}$$

Once property (9) is established, the theorem is in hand; for if $W_t(\phi)$ satisfies (6), it will then follow from the sufficiency of (8) that $V_t(W_t(\phi))$ solves problem (2).

In order to establish (9), we examine the quantity

$$g_{\ell}(t;H_{\ell})|_x = DV_t(\zeta)\sum_{\nu=0}^{\infty}\frac{(-t)^{\nu}}{\nu!}\left[ad^{\nu}_{H_0}H_{\ell}\right]\zeta|_x - H_{\ell}V_t(\zeta)|_x. \tag{10}$$

With ζ an element of the allowed domain, the maps $t \to V_t(\zeta)|_x$ and $t \to H_{\ell}V_t(\zeta)|_x$ are complex analytic by our hypotheses, as is the map $t \to DV_t(\zeta)|_x$. Consequently, the right-hand side of (10) is complex analytic in t, for all ζ and for all x. Therefore it is legitimate to evaluate $g_{\ell}(t;H_{\ell})|_x$ be means of its Taylor expansion in t.

To begin with, we know $g_{\ell}(0;H_{\ell})|_x = 0$, because $DV_t(\zeta) \to id$ in the strong operator topology as $t \to 0^+$, and $V_0(\zeta) = \zeta$. Next, consider that

$$\begin{aligned}\frac{d}{dt}DV_t(\zeta) = D\frac{\partial}{\partial t}V_t(\zeta) = D_{\zeta}[H_0V_t(\zeta)] &= D_{\zeta}[H_0(V_t(\zeta))]\\ &= (DH_0(V_t(\zeta)))(DV_t(\zeta)) \quad .\end{aligned}$$

(The differentials in the first line are all with respect to ζ, as is indicated explicitly in places where confusion might arise. The differential $D_{\zeta}[H_0(V_t(\zeta))]$ is computed as the product of the differential of the mapping $V_t(\zeta) \to H_0(V_t(\zeta))$ and the differential of the initial mapping $\zeta \to V_t(\zeta)$.) In similar vein,

$$\frac{d}{dt}[H_\ell V_t(\zeta)] = \frac{d}{dt}[H_\ell(V_t(\zeta))] = DH_\ell(V_t(\zeta))H_0V_t(\zeta) \quad .$$

Using these last two relations, we may obtain (with the dot indicating time derivative)

$$\begin{aligned}\dot{g}_\ell(t;H_\ell) &= DH_0(V_t(\zeta))DV_t(\zeta)\sum_{\nu=0}^{\infty}\frac{(-t)^\nu}{\nu!}[ad^\nu_{H_0}H_\ell]\zeta \\ &\quad - DV_t(\zeta)\sum_{\nu=0}^{\infty}\frac{(-t)^\nu}{\nu!}[ad^{\nu+1}_{H_0}H_\ell]\zeta - \frac{d}{dt}[H_\ell V_t(\zeta)] \\ &= DH_0(V_t(\zeta))[DV_t(\zeta)\sum_{\nu=0}^{\infty}\frac{(-t)^\nu}{\nu!}[ad^\nu_{H_0}H_\ell]\zeta - H_\ell V_t(\zeta)] \\ &\quad + [DH_0(V_t(\zeta))H_\ell V_t(\zeta) - DH_\ell(V_t(\zeta))H_0V_t(\zeta)] \\ &\quad - DV_t(\zeta)\sum_{\nu=0}^{\infty}\frac{(-t)^\nu}{\nu!}[ad^{\nu+1}_{H_0}H_\ell]\zeta \\ &= DH_0(V_t(\zeta))g_\ell(t;H_\ell) - g_\ell(t;ad_{H_0}H_\ell) \quad . \qquad (11)\end{aligned}$$

But we know, from previous argument or its extension, that $g_\ell(t;H_\ell)|_x$ and $g_\ell(t;ad_{H_0}H_\ell)|_x$ tend to zero as $t \to 0^+$; it follows that $\dot{g}_\ell(0;H_\ell)|_x = 0$ for all ζ and for all x.

The pattern is now set for an inductive construction of successive time derivatives of $g(t;H_\ell)$. In particular, based on the above results we may form

$$\begin{aligned}\ddot{g}_\ell(t;H_\ell) &= \frac{d}{dt}[DH_0(V_t(\zeta))g_\ell(t;H_\ell)] + DH_0(V_t(\zeta))\dot{g}_\ell(t;H_\ell) \\ &\quad - DH_0(V_t(\zeta))\dot{g}_\ell(t;ad_{H_0}H_\ell) + g_\ell(t;ad^2_{H_0}H_\ell) \quad ,\end{aligned}$$

and it follows that $\ddot{g}_\ell(t;H_\ell)|_x \to 0$ as $t \to 0^+$. Continuing the process indefinitely, we arrive at the result that at $t = 0$ all the time derivatives of $g_\ell(t;H_\ell)|_x$ vanish, to arbitrarily high order. Thus $g_\ell(t;H_\ell)|_x$ is identically 0, $V\zeta$, Vx, i.e.,

$$DV_t(\zeta)\sum_{\nu=0}^{\infty}\frac{(-t)^\nu}{\nu!}[ad^\nu_{H_0}H_\ell]|_x = H_\ell V_t(\zeta)|_x \quad ,$$

$\ell = 1, \ldots, r$. The desired property (9) ensues upon multiplying this equality by $u_\ell(t)$ and summing over ℓ.

Corollary 1. Same as Theorem 1, except that "complex analytic" is everywhere to be replaced by "real analytic". (See Ref. 18)

Proof. Direct observation.

IV. LOCAL CONTROLLABILITY ALONG A REFERENCE HOMOGENEOUS SOLUTION

Definition 2. The system (2) is said to be locally controllable along the solution $\eta_t = V_t(\phi)$ of the control-free problem (5) on the manifold $M \subset H$ if, for small $t > 0$, there exists a set of $u_\ell(t)$, $\ell = 1, \ldots, r$, such that the solution ψ_t of (2) can be controlled to a neighborhood of η_t on M. The precise meaning of the last phrase is that ψ_t can be steered into any direction of the tangent space TM_{η_t} of M at the point $\eta_t = V_t(\phi) \in M$, $\forall\, \phi \in M$.

We may now formulate the central result of this section.

Theorem 2. Assume that the homogeneous solution of system (2), i.e., the solution $\eta_t = V_t(\phi)$ of the uncontrolled system (5), satisfies the hypotheses (i) and (ii) of Theorem 1 for ζ (and specifically ϕ) on a finite-dimensional submanifold M, $M \subset D \subset H$, dim $M = m$. Assume further that there exist integers $\nu_{\ell_{j_\ell}}$ (with $\ell = 1, \ldots, r$ and $j_\ell = 1, \ldots, k_\ell < \infty$, and $0 \leq \nu_{\ell 1} < \nu_{\ell 2} < \ldots < \nu_{\ell k_\ell}$) such that the set $\{[\mathrm{ad}_{H_0}^{\nu_{\ell j_\ell}} H_\ell]\phi\}$ spans the tangent space TM_{η_t} of M at $\eta_t = V_t(\phi)$ for all $\phi \in M$. It follows that system (2) is locally controllable along η_t on M. (Cf. Theorem 2, Ref. 18.)

Proof. If the functions $u_{\ell j_\ell}(t)$, where $\ell = 1, \ldots, r$ and $j_\ell = 1, \ldots, k_\ell$, qualify as admissible controls (real, analytic, bounded functions of t), then so do the finite linear combinations

$$u_\ell^{a_\ell}(t) \sum_{j_\ell=1}^{k_\ell} a_{\ell j_\ell} u_{\ell j_\ell}(t), \qquad \ell = 1, \ldots, r ,$$

wherein the real coefficients $a_{\ell 1}, \ldots, a_{\ell k_\ell}$, are chosen (for convenience) to obey

$$\sum_{j_\ell=1}^{k_\ell} |a_{\ell j_\ell}| = 1 \quad , \ell = 1, \ldots, r .$$

Let us abbreviate the set $\{a_{\ell j_\ell}\}$ simply as $\underline{a}$. By generalized decomposition in the multi-input, complex case of the preceding subsection (i.e., by virtue of Theorem 1, the solution of problem (2), with the $u_\ell^{a_\ell}$ as controls, is given by $\psi_t^a = V_t(W_t^a(\phi))$. The solution $\xi_t^a = W_t^a(\phi)$ of the boundary value problem (6), restated for the controls $u_\ell^{a_\ell}$, evidently obeys the integral equation

$$W_t^a(\phi) = \phi + \sum_{\ell=1}^{r} \int_0^t u_\ell^{a_\ell}(t) \sum_{\nu=0}^{\infty} \frac{(-s)^\nu}{\nu!}\, ds \left[\mathrm{ad}_{H_0}^{\nu} H_\ell\right] W_t^a(\phi) \quad .$$

Thus

$$\frac{\partial}{\partial a_{\ell j_\ell}} V_t(W_t^a(\phi)) \mid_{a=0} = DV_t(W_t^a(\phi)) \frac{\partial}{\partial a_{\ell j_\ell}} W_t^a(\phi) \mid_{a=0}$$

$$= \sum_{\nu=0}^{\infty} \left(\int_0^t u_{\ell j_\ell}(s) \frac{(-s)^\nu}{\nu!} ds\right) DV_t(\phi)\left[ad_{H_0}^{\nu} H_\ell\right]\phi \ , \tag{12}$$

where $a=0$ means all of the $a_{\ell j_\ell}$ are zero. By assumption, we can find a set of integral (or zero) powers $\nu_{\ell j_\ell}$, where $\ell = 1, \ldots, r$, $j_\ell = 1, \ldots, k_\ell$, $0 \leq \nu_{\ell 1} < \nu_{\ell 2} < \ldots < \nu_{\ell k_\ell}$, and $\nu_{max} = \max(\nu_{\ell j_\ell}) < \infty$, such that the set $\{[ad_{H_0}^{\nu_{\ell 1}} H_\ell]\phi, \ldots, [ad_{H_0}^{\nu_{\ell k_\ell}} H_\ell]\ \phi, \ \ell = 1, \ldots, r\}$ spans TM_{η_t}. Then, since (also by assumption) $DV_t(\phi) \to id$ strongly as $t \to 0^+$, there must exist a time $t_1 > 0$ such that the set

$$\left\{DV_t(\phi)\left[ad_{H_0}^{\nu_{\ell 1}} H_\ell\right]\phi, \ \ldots, \ DV_t(\phi)\left[ad_{H_0}^{\nu_{\ell k_\ell}} H_\ell\right]\phi\right\}$$

spans TM_{η_t}, over the time interval $0 \leq t \leq t_1$.

We now proceed to make a judicious choice of the original functions $u_{\ell j_\ell}(t)$ involved in (12). One can realize admissible controls $\bar{u}_{\ell j_\ell}(t)$ obeying the conditions

$$\int_0^{t_1} \bar{u}_{\ell j_\ell}(s) \frac{(-s)^\nu}{\nu!} ds = \begin{cases} 0, & \text{for } \nu \neq \nu_{\ell j_\ell} \ , \ \ 0 \leq \nu \leq \nu_{max} + 1 \ , \\ c_{\ell j_\ell} \neq 0 \ , & \text{for } \nu = \nu_{\ell j_\ell} \ , \end{cases} \tag{13}$$

where $\ell = 1, \ldots, r$, $j_\ell = 1, \ldots, k_\ell$, and the $c_{\ell j_\ell}$ are real constants. The connection between the $\bar{u}_{\ell j_\ell}$ and the $u_{\ell j_\ell}$ will be specified shortly. The power ν being integral, inversion of relations (13) is in effect just a classical finite-moments problem. (Note that in the upper range $\nu > \nu_{max} + 1$, we have

$$\int_0^{t_1} \bar{u}_{\ell j_\ell}(s) \frac{(-s)^\nu}{\nu!} ds = 0\left(t_1^{\nu_{max} + 3}\right) \ ,$$

since $|\bar{u}_{\ell j_\ell}|$ is by assumption bounded. This implies that the higher moments not specified by (13) will be negligible.)

With t in the interval $[0,t_1]$, we now carry out the change of variable $s = t_1 h/t$ in the integral on the left of (13):

$$\int_0^{t_1} \bar{u}_{\ell j_\ell}(s) \frac{(-s)^\nu}{\nu!} ds = \left(\frac{t_1}{t}\right)^{\nu+1} \int_0^t u_{\ell j_\ell}(t_1 h/t) \frac{(-h)^\nu}{\nu!} dh .$$

Hence

$$\int_0^t \bar{u}_{\ell j_\ell}(t_1 h/t) \frac{(-h)^\nu}{\nu!} dh = \begin{cases} 0, & \text{for } \nu \neq \nu_{\ell j_\ell},\ 0 \le \nu \le \nu_{\max} + 1 , \\ \left(\frac{t}{t_1}\right)^{\nu+1} c_{\ell j_\ell} , & \text{for } \nu = \nu_{\ell j_\ell} , \\ 0\left(t^{\nu_{\max} + 3}\right) , & \text{for } \nu > \nu_{\max} + 1 . \end{cases}$$

Setting $u_{\ell j_\ell}(s) = \bar{u}_{\ell j_\ell}(t_1 s/t)$ in (12), we arrive finally at the result

$$\frac{\partial}{\partial a_{\ell j_\ell}} V_t(W_t^a(\phi)) \,|_{a=0} = c_{\ell j_\ell}\left(\frac{t}{t_1}\right)^{\nu_{\ell j_\ell} + 1} DV_t(\phi) \left[ad_{H_0}^{\nu_{\ell j_\ell}} H_\ell\right]\phi + 0 \left(t^{\nu_{\max} + 2}\right) ,$$

where, for $t < t_1$, the last term can be neglected, t_1 being small. Consequently the set $\{\partial V_t(W_t^a(\phi))/\partial a_{\ell j_\ell}, \ell = 1, \ldots, r, j_\ell = 1, \ldots, k_\ell\}$ spans TM_{η_t} for t in the interval $[0,t_1]$, where t_1 has been chosen above. This means that we have been able to choose the controls so that, for small $t > 0$, the state defined by system (2) can be steered into any direction of the tangent space on M at the point $\eta_t = V_t(\phi)$. Then by definition the system is locally controllable along the reference solution $V_t(\phi)$, for all $\phi \in M$.

Remark 3. Theorems 1 and 2 remain true as stated if the H_k, $k = 0, \ldots, r$, are not skew-Hermitian.

Example 1. The theorems of the present paper are aimed at an infinite-dimensional space of quantum states. However, the results obtained herein are still valid (with trivial alterations) for a finite-dimensional state space. As pointed out in Remark 3, from a mathematical standpoint we may also dispense with the assumption that the generators $H_0, \ldots, H_r$ are skew-Hermitian.

For example, consider a nonlinear control system on R^m, $m < \infty$, defined by

$$\frac{d}{dt} \chi(t) = A(\chi(t)) + u(t)B(\chi(t)) \quad , \chi(0) = \chi_0 , \tag{14}$$

where A and B are real analytic vector fields corresponding to nonlinear operators of the sort introduced in Section II. Then, as argued in Ref. 18, a sufficient condition for local controllability along the homogeneous ($u \equiv 0$) solution of (14) is $\text{rank}\{[ad_A^\nu B]\chi_0, \nu = 0,1,2,\ldots,\infty\} = m$. This is precisely the condition which would enter the finite-state-space version of Theorem 2. Problem (14) does not strictly refer to a quantum-mechanical system; its study is, nevertheless, illuminating.

While surely of high interest, the identification and analysis of "non-trivial" examples of the utility of Theorem 2, meaning examples concerned with novel quantum control systems characterized by nonlinear generators, exceeds the scope of the present work.

V. SUMMARY AND OUTLOOK

It has been our aim to augment the foundation for the concept of controllability of quantum-mechanical systems [12]. In the generalized, nonlinear formulation of the quantum control problem, we were able to determine conditions for the property of local controllability along a homogeneous (i.e., control-free) solution, without having to refer to the existence of an analytic domain which was assumed in the global analysis of [12]. (Our treatment of this case amounts to an extension of Hermes' work [18] to a multi-input, complex-state problem.) From the results obtained herein on the controllability of the solution of nonlinear Schrodinger equations, one may regain, upon appropriate specialization or adaptation, certain well-known systems-theoretic results in finite-dimensional state space (see, in particular, Refs. 13-18).

Clearly, only a modest beginning has been made toward achieving the larger goal of a comprehensive theory of quantum control. The following problems, among others, await concerted effort:

(i) Adaptation of the notions of observability, identification, realization, and feedback to the quantum context [27].

(ii) Study of a controlled version of the Schrodinger equation for the time evolution of the density operator, [19] so as to extend control theory to the realm of quantum statistical mechanics.

It is evident that powerful mathematical techniques must be invoked to carry through this program; moreover, one must confront the profound conceptual obstacles intrinsic to the quantum measurement process [25,28,29].

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under Grant Nos. DMR-8519077 and ECS-8515899.

REFERENCES

[1] Tarn, T.J., G.M. Huang and J.W. Clark: Modelling of quantum mechanical control systems, Mathematical Modelling, 1 (1980), 109-121.

[2] van der Schaft, Aryan J.: Hamiltonian and quantum mechanical control systems, in: Proceedings of the 4th International Seminar on Mathematical Theory of Dynamical Systems and Microphysics, Udine (Ed. A. Blaquiere, S. Diner and G. Lochak), Springer-Verlag, Wien-New York 1987.

[3] Clark, J.W. and T.J. Tarn: Quantum nondemolition filtering, in: Proceedings of the 4th International Seminar on Mathematical Theory of Dynamical Systems and Microphysics, Udine(Ed. A. Blaquiere, S. Diner and G. Lochak), Springer-Verlag, Wien-New York 1987.

[4] Tarn, T.J., J.W. Clark, C.K. Ong and G.M. Huang: Continuous-time quantum mechanical filter, in: Proceedings of the Joint Workshop on Feedback and Synthesis of Linear and Nonlinear Systems, Bielefeld and Rome (Ed. D. Hinrichsen and A. Isidori), Springer-Verlag, Berlin 1982.

[5] Clark, J.W., C.K. Ong, T.J. Tarn and G.M. Huang: Quantum nondemolition filters, Mathematical Systems Theory, 18(1985), 33-35.

[6] Ong, C.K., G.M. Huang, T.J. Tarn and J.W. Clark: Invertibility of quantum-mechanical control systems, Mathematical Systems Theory, 17(1984), 335-350.

[7] Belavkin, Viacheslav: Non-demolition measurement and control in quantum dynamical systems, in : Proceedings of the 4th International Seminar on Mathematical Theory of Dynamical Systems and Microphysics, Udine (Ed. A. Blaquiere, S. Diner and G. Lochak), Springer-Verlag, Wien-New York 1987.

[8] Peirce, A.P., M.A. Dahleh and H. Rabitz: Optimal control of quantum mechanical systems: existence, numerical approximations, and applications, to appear in: Proceedings of the IEEE International Conference: Control 88, University of Oxford, UK, April (1988).

[9] Butkovskiy, A.G. and Yu. I. Samoilenko, Control of quantum systems, automation and remote control, No. 4, April (1979), 485-502; Control of quantum systems, automation and remote control, No. 5 May (1979), 629-645.

[10] Butkovskiy, A.G. and Ye. I. Pustil'nykova: The method of seeking finite control for quantum mechanical processes, in: Proceedings of the 4th International Seminar on Mathematical Theory of Dynamical Systems and Microphysics, Udine (Ed. A. Blaquiere, S. Diner and G. Lochak), Springer-Verlag, Wien-New York 1987.

[11] Butkovskiy, A.G. and Yu. I. Samoilenko: Controllability of quantum mechanical systems, Dokl, Akad. Nauk SSSR 250, 51 (1980) [Sov, Phys, Dokl, 25, 22 (1980)].

[12] Huang, G.M., T.J. Tarn and J.W. Clark: On the controllability of quantum mechanical systems, J. Math. Phys. 24 (1983) 2608-2618.

[13] Sussmann, H. and V. Jurdjevic: Controllability of non-linear systems, Journal of Differential Equations, 12 (1962), 95-116.

[14] Jurdjevic, V. and H. Sussmann: Control systems on lie groups, Journal of Differential Equations, 12 (1972), 313-329.

[15] Krener, Arthur J.: A Generalization of chow's theorem and the bang-bang theorem to nonlinear control problems, SIAM Journal of Control, Vol. 12, No. 1, Feb. (1974).

[16] Brockett, Roger W.: Nonlinear systems and differential geometry: Proceedings of IEEE, Vol. 64, No. 1, Jan (1976).

[17] Kunita, Hiroshi: On the controllability of nonlinear systems, with applications of polynomial systems, Applied Mathematics and Optimization (1976), 89-99.

[18] Hermes, H.: Local controllability of observables in finite and infinite dimensional nonlinear control systems, Applied Mathematics and Optimization, 5, (1979), 117-125.

[19] Messiah, A.: Quantum mechanics, Vols. I and II, Wiley, New York (1961).

[20] Beals, R. and C. Feffermann: On Local solvability of linear partial differential equations, Annals of Mathematics, 97 (1973), 483-498.

[21] Brown, G.E.: Unified theory of nuclear models and forces, North-Holland, Amsterdam, (1971).

[22] Santilli, R.M.: Need of subjecting to an experimental verification the validity within a hadron of Einstein's special relativity and Pauli's exclusion principle, Hadronic Journal, 1, (1978), 574-901.

[23] Abraham, R., and J.E. Marsden: Foundations of mechanics, 2nd ed. Benjamin, Reading, (1978).

[24] Wigner, E.P.: The Scientist speculates, I.J. Good, Ed. W. Heinemann, London, (1961).

[25] d'Espagnat, B.: Conceptual foundations of quantum mechanics, Benjamin, Reading, (1976).

[26] Hermes, H.: Controllability of nonlinear delay differential equations, Nonlinear Analysis, Theory, Methods and Applications, 3 (1979).

[27] Kailath, T.: Linear systems, Prentice-Hall, Inc., Englewood Cliffs, (1980).

[28] Helstrom, C.W.: Quantum detection and estimation theory, Academic Press, New Yok, (1976).

[29] Ilic, D.: D. Sc. Dissertation, Washington University (1978), unpublished.

FROM TWO STOCHASTIC OPTIMAL CONTROL PROBLEMS TO THE SCHRODINGER EQUATION

K. Kime
Department of Chemistry, Princeton University
Princeton, NJ 08544 (USA)

A. Blaquiere
Universite Paris 7, Laboratoire d'Automatique Theorique
Paris (FRANCE)

1. Introduction

In recent years, interest has developed in the connections between stochastic control theory, dynamic programming and quantum mechanics [1-4, 7, 12, 13] and (related) variational approaches [9, 11, 14, 15] to Nelson's stochastic mechanics [10]. In this paper, we will start by considering two stochastic optimal control problems, one "forward" in time, one "backward" in time. We show that, if there are solutions to the extended Hamilton-Jacobi equations associated with the control problems, then there is a solution of a Schrödinger equation and conversely, if there is a sufficiently well-behaved solution to a Schrödinger equation, there are solutions to a pair of extended H-J equations. We note the connection between the H-J equations and the main dynamical equations of Nelson's stochastic mechanics. The H-J equations are equivalent to a pair of inhomogeneous "backward" and "forward" heat equations via a well-known exponential transformation. One may thus pass from these to a Schrödinger equation (and back).

2. Definitions and Notations

We assume a given underlying probability space (Ω,F,P). E^n denotes n-dimensional Euclidean space, (t_0,t_1) an interval in E^1. S denotes $(t_0,t_1) \times E^n$; $\bar{S} = [t_0,t_1] \times E^n$. Definitions of stochastic process, Brownian motion will be taken from [6] as will other elements of our framework which will be noted below.

A solution of a stochastic differential equations

$$d\xi = b(t,\xi(t))dt + \sigma(t,\xi(t))dw \qquad (2.1)$$

with initial data $\xi(s) = y$ is to be interpreted as in [6] as a solution of the integral equation

$$\xi(t) = \xi(s) + \int_s^t b(r,\xi(r))dr + \int_s^t \sigma(r,\xi(r))dw(r) \qquad (2.2)$$

Here, w is standard Brownian motion of dimension n. With the vector notation $\xi=(\xi_1,\ldots\xi_n)$, $b=(b_1\ldots b_n)$, we have

$$d\xi_i = b_i(t,\xi(t))dt + \sum_{\ell=1}^{n} \sigma_{i\ell}(t,\xi(t))dw_\ell \qquad i=1,\ldots n$$

The notation $C_p^{1,2}(S)$ denotes the class of functions ψ in $C^{1,2}(S)$ (meaning C^1 in t, C^2 in x) which satisfy $|\psi(t,x)| \le D(1+|x|^k)$ for some constants D,k, when $(t,x) \in S$.

3. Two Stochastic Optimal Control Problems

We consider first a "forward" stochastic optimal control problem, Problem 1, in 3.1, then the symmetric "backward" problem, Problem 2, in 3.2. The controls v and $\tilde{v}$ will take values in E^n.

3.1 Problem 1

Consider the stochastic differential equation

$$d\xi = v(t,\xi(t))dt + \sigma dw \tag{3.1.1}$$

with initial data $\xi(s) = x \in E^n$, at time $s \in (t_0,t_1)$. Here, w is a standard n-dimensional Brownian motion, and

$$\sigma_{ij} = \sqrt{2D}\,\delta_{ij}$$

where δ is the Kronecker delta, and D is a positive constant. We assume that v belongs to a class of admissible control functions defined as follows:

Definition 3.1.A [6]. A feedback control law v (the term feedback refers to the fact that the control is a function of the state $\xi(t)$) is admissible if v is a Borel measurable function from $\bar{S}$ into E^n, such that

(a) For each (s,x), $t_0 \le s \le t_1$, there exists a Brownian motion w such that (3.1.1) with initial data $\xi(s) = x$ has a solution ξ, unique in probability law

(b) For each $k > 0$, $E_{sx}|\xi(t)|^k$ is bounded for $s \le t \le t_1$, and

$$E_{sx}\int_s^{t_1} |v(t,\xi(t))|^k\, dt < \infty$$

(the bound may depend on (s,x)). The subscript sx refers to the fact that $\xi(s) = x$.

Either of the following conditions are sufficient for the admissibility of v:

(i) For some constant M_1, $|v(t,y)| \le M_1(1+|y|)$ for all $(t,y) \in \bar{S}$. Moreover, for any bounded Borel set $B \subset E^n$ and $t_0 < t' < t_1$, there exists a constant K_1 such that, for all $x,y \in B$ and $t_0 \le t \le t'$,

$$|v(t,x) - v(t,y)| \le K_1|x-y|$$

(K_1 may depend on B,t'; and both M_1, K_1 may depend on v).

(ii) v satisfies a Lipschitz condition on $\bar{S}$. Further, if (i) or (ii) holds, the Brownian motion w can be specified in advance, which is the case in Problem 1.

Now, for $(t,x) \in \bar{S}$ and $v \in E^n$, let

$$L(t,x,v) = \frac{1}{2} mv^2 + Q(t,x) \tag{3.1.2}$$

where Q is continuous on $\bar{S}$, and let $W_1 : E^n \to R_+$ (R_+ denoting non-negative real numbers) be continuous and assume

$$\begin{aligned} |Q(t,x)| &\leq C(1 + |x|)^k \\ W_1(x) &\leq C(1 + |x|)^k \end{aligned} \tag{3.1.3}$$

for some constants C,k.

We define a <u>cost function</u>

$$J(s,x,v) = E_{sx} \left\{ \int_s^{t_1} L(t,\xi(t),v(t,\xi(t)))\, dt + W_1(\xi(t_1)) \right\} .$$

The conditions on Q and W_1 ensure that J is finite.

Now let the optimal control problem be as follows: Find an admissible feedback control $v*$, among all admissible feedback controls, which minimizes $J(s,x,v)$. The following Verification Theorem gives sufficient conditions for the existence of a minimizing v^*.

<u>Theorem 3.1.B [6].</u> Let $W(s,x)$ be a solution of the dynamic programming equation

$$0 = \frac{\partial W}{\partial s} + \min_{v \in E^n} \left[D\Delta W + \sum_{i=1}^{n} v_i \frac{\partial W}{\partial x_i} + \frac{1}{2} mv^2 + Q(s,x) \right] \tag{3.1.4}$$

$$(s,x) \in S \quad ,$$

with boundary data

$$W(t_1,x) = W_1(x), \qquad x \in E^n, \tag{3.1.5}$$

such that W is in $C_p^{1,2}(S)$ and continuous on $\bar{S}$. Then,

(a) $W(s,x) \leq J(s,x,v)$ for any admissible feedback control v and any initial data $(s,x) \in S$.

(b) If v^* is an admissible feedback control such that

$$D\Delta W + \sum_{i=1}^{n} v_i^*(s,x) \frac{\partial W}{\partial x_i} + \frac{1}{2} m(v^*(s,x))^2 + Q(s,x)$$

$$= \min_{v \in E^n} \left[D\Delta W + \sum_{i=1}^{n} v_i \frac{\partial W}{\partial x_i} + \frac{1}{2} mv^2 + Q(s,x) \right] \tag{3.1.4}$$

for all $(s,x) \in S$, then $W(s,x) = J(s,x,v^*)$ for all $(s,x) \in S$.

Thus v^* is optimal

Now let us assume that there exists a W satisfying the hypotheses of the Verification Theorem, and an optimal control v^*. Then, since the controls take values in E^n, which is open

$$mv^* = - \text{grad } W \text{ for all } (s,x) \in S \tag{3.1.6}$$

and

$$\frac{\partial W}{\partial s} = - D\Delta W + \frac{1}{2m} (\text{grad } W)^2 - Q \tag{3.1.7}$$

for all $(s,x) \in S$. Equation (3.1.7) is analogous to the Hamilton-Jacobi equation of classical mechanics; we shall refer to it as an extended Hamilton-Jacobi equation.

3.2 Problem 2

Now let us introduce another type of admissibility for a feedback control function as follows:

Definition 3.2.A A feedback control law $\bar{v}$ is backward admissible if $\bar{v}$ is such that

$$\bar{v}(\tau,x) = - \hat{v}(t_0+t_1-\tau,x) \text{ for all } (\tau,x) \in \bar{S}, \text{ and}$$

$\hat{v}$ is an admissible feedback control law.

We consider the stochastic differential equation

$$d\eta = \bar{v}(\tau,\eta(\tau))\, d\tau + \sqrt{2D}\ \ d\bar{w} \tag{3.2.1}$$

where $\bar{v}$ is a backward admissible feedback control law, and

$$\bar{w}(\tau) = w(t_0+t_1-\tau).$$

We say that η is a solution to (3.2.1) with terminal data $\eta(\sigma) = y \in E^n$, with $t_0 \le \tau < \sigma \le t_1$, if η satisfies the integral equation

$$\eta(\tau) = \eta(\sigma) - \int_\tau^\sigma \bar{v}(r,\eta(r))\, dr - \int_\tau^\sigma \sqrt{2D}\, d\bar{w}(r). \tag{3.2.2}$$

By making the change of variable

$$\tau = t_0 + t_1 - t,$$
$$\sigma = t_0 + t_1 - s,$$
$$\ell = t_0 + t_1 - r$$

(3.2.2) becomes

$$\eta(t_0+t_1-t) = \eta(t_0+t_1-s) - \int_t^s \bar{v}(t_0+t_1-\ell,\ \eta(t_0+t_1-\ell))\ (-d\ell)$$
$$- \int_t^s \sqrt{2D}\ d\bar{w}\ (t_0+t_1-\ell) \tag{3.2.3}$$

Define $\hat{\eta}(t) = \eta(t_0+t_1-t)$. Now (3.2.3) becomes

$$\hat{\eta}(t) = \hat{\eta}(s) + \int_s^t \hat{v}\,(\ell,\hat{\eta}(\ell))\,d\ell + \int_s^t \sqrt{2D}\,d\bar{w}(\ell) \quad , \tag{3.2.4}$$

and we have

$$\hat{\eta}(s) = y \quad . \tag{3.2.5}$$

We now define

$$\hat{J}(s,y,\hat{v}) = E_{sy}\left\{\int_s^{t_1} [\tfrac{m}{2}\,\left(\hat{v}(\ell,\hat{\eta}(\ell))\right)^2 + \hat{Q}(\ell,\hat{\eta}(\ell))]d\ell + \bar{W}_0(\hat{\eta}(t_1))\right\}$$

$$= E_{\sigma y}\left\{\int_{t_0}^{\sigma} [\tfrac{m}{2}\,\left(\bar{v}(r,\eta(r))\right)^2 + Q(r,\eta(r))]dr + \bar{W}_0\,(\eta(t_0))\right\}$$

$$= \bar{J}(\sigma,y,\bar{v}) \quad . \tag{3.2.6}$$

Here Q is the same as in Problem 1, $\bar{W}_0: E^n \to R_+$ is continuous and $\bar{W}_0(y) \le C(1+|y|)^k$, (C,k as in (3.1.3)). Thus, $\hat{Q}(\ell,\hat{\eta}(\ell)) = Q(t_0+t_1-\ell,\ \eta(t_0+t_1-\ell))$.

We now consider, as in Problem 1, the problem of minimizing (3.2.6). For given terminal data $y \in E^n$ at time $\sigma \in (t_0,t_1]$, we shall say that $\bar{v}_*$ is <u>backward optimal</u> if $\bar{v}_*$ is backward admissible, and

$$\bar{J}\,(\sigma,y,\bar{v}) \ge \bar{J}\,(\sigma,y,\bar{v}_*)$$

for all backward admissible $\bar{v}$.

In view of (3.2.4) - (3.2.6), we have the following version of the Verification Theorem:

<u>Theorem 3.2.B</u> Let $\hat{W}$ be a solution of the dynamic programming equation

$$0 = \frac{\partial\hat{W}}{\partial s} + \min_{\hat{\upsilon}\in E^n}\left[D\Delta\hat{W} + \sum_{i=1}^n \hat{\upsilon}_i\frac{\partial\hat{W}}{\partial y_i} + \frac{1}{2}m\hat{\upsilon}^2 + \hat{Q}(s,y)\right] \tag{3.2.7}$$

$$(s,y) \in S \quad ,$$

with $\hat{W}(t_1,y) = \bar{W}_0(y)$, $y \in E^n$, such that $\hat{W}$ is in $C_p^{1,2}(S)$ and continuous on $\bar{S}$. Then:

(a) $\hat{W}(s,y) \le \hat{J}\,(s,y,\hat{v})$ for any admissible feedback control $\hat{v}$ and any initial data $(s,y) \in S$.

(b) If $\hat{v}^*$ is an admissible feedback control such that

$$D\Delta\hat{W} + \sum_{i=1}^n \hat{v}_i^*\,(s,y)\,\frac{\partial\bar{W}}{\partial y_i} + \frac{1}{2}m(\hat{v}^*(s,y))^2 + \hat{Q}(s,y) =$$

$$\min_{\hat{\upsilon}\in E^n}\left[D\Delta\hat{W} + \sum_{i=1}^n \hat{\upsilon}_i\frac{\partial\hat{W}}{\partial y_i} + \frac{1}{2}m\hat{\upsilon}^2 + \hat{Q}(s,y)\right] \tag{3.2.8}$$

for all $(s,y) \epsilon S$, then $\hat{W}(s,y) = \hat{J}(s,y,\hat{v}*)$ for all $(s,y) \epsilon S$; $\hat{v}^*$ is optimal

Now suppose there exists a function $\hat{W}$ satisfying these hypotheses, and an optimal control $\hat{v}^*$. Define

$$\bar{W}(\sigma,y) = \hat{W}(t_0+t_1-\sigma,y) \ , \ t_0 < \sigma \le t_1 \ .$$

Then $\bar{W}(t_0,y) = \hat{W}(t_1 y)$ and $\frac{\partial\bar{W}}{\partial\sigma} = - \frac{\partial\hat{W}}{\partial s}$.

We define

$$\bar{v}_*(\sigma,y) = \bar{v}_*(t_0+t_1-s,y) = - \hat{v}^* \ (s,y).$$

Now we have

$$0 = - \frac{\partial\bar{W}}{\partial\sigma} + D\Delta\bar{W} - \sum_{i=1}^{n} (\bar{v}_{*}i(\sigma,y)) \frac{\partial\bar{W}}{\partial y_i} + \frac{1}{2} m(\bar{v}_*(\sigma,y))^2 + Q(\sigma,y) \tag{3.2.9}$$

and, as in Problem 1,

$$m\bar{v}_* = \text{grad } \bar{W} \tag{3.2.10}$$

$$\frac{\partial\bar{W}}{\partial\sigma} = D\Delta\bar{W} - \frac{1}{2m} (\text{grad } \bar{W})^2 + Q \qquad \text{on } S. \tag{3.2.11}$$

We have let

$$\hat{v}^*(s,y) = - \bar{v}_*(t_0+t_1-s,y) \ , \quad y \epsilon E^n.$$

From (3.2.6) we have

$$\bar{J}(\sigma,y,\bar{v}_*) = \hat{J}(s,y,\hat{v}^*). \tag{3.2.12}$$

If the Verification Theorem 3.2.B is satisfied, then $\hat{v}*$ is optimal; that is

$$\hat{J}(s,y,\hat{v}) \ge \hat{J}(s,y,\hat{v}^*). \tag{3.2.13}$$

From (3.2.6) and (3.2.11), (3.2.12) implies

$$\bar{J}(\sigma,y,\bar{v}) \ge \bar{J} \ (\sigma,y,\bar{v}_*). \tag{3.2.14}$$

for all backward admissible $\bar{v}$.

Therefore, if $\hat{v}^*$ is an optimal control in the sense of Theorem 3.2.A, then $\bar{v}_*$ is a backward optimal control for Problem 2, and the converse is also true.

4. Extended Hamilton-Jacobi Equations, the Schrödinger Equation and Inhomogeneous Backward and Forward Heat Equations.

4.1 Extended Hamilton-Jacobi Equations and the Schrödinger Equation

We have seen, that if there exist W, $\bar{W}$, v^*, $\bar{v}^*$ satisfying the conditions of the Verification Theorems, then W is a solution of the equation

$$\frac{\partial G}{\partial t}(t,x) - \frac{1}{2m}(\mathrm{grad}\ G(t,x))^2 + D\Delta G(t,x) + Q(t,x) = 0 \qquad (t,x) \in S \tag{4.1.1}$$

with

$$G(t_1,x) = W_1(x), \tag{4.1.2}$$

and $\bar{W}$ is a solution of the equation

$$\frac{\partial \bar{G}}{\partial t}(t,x) + \frac{1}{2m}(\mathrm{grad}\ \bar{G}(t,x))^2 - D\Delta\bar{G}(t,x) - Q(t,x) = 0 \qquad (t,x) \in S \tag{4.1.3}$$

with

$$\bar{G}(t_0,x) = \bar{W}_0(x). \tag{4.1.4}$$

We now show that, when there are solutions G, $\bar{G}$ of (4.1.1), (4.1.3), then there are solutions of a Schrödinger equation. From now on D shall denote $\hbar/2m$.

Proceeding as in [4], with $G^* = \frac{\bar{G}+G}{2}$, $H^* = \frac{\bar{G}-G}{2}$, we have

$$\frac{\partial}{\partial t}(G^*-H^*) - \frac{1}{2m}(\mathrm{grad}(G^*-H^*))^2 + D\Delta(G^*-H^*) + Q = 0 \tag{4.1.5}$$

$$\frac{\partial}{\partial t}(G^*+H^*) + \frac{1}{2m}(\mathrm{grad}(G^*+H^*))^2 + D\Delta(G^*+H^*) - Q = 0 \tag{4.1.6}$$

Adding and subtracting (4.1.5), (4.1.6) gives

$$\frac{\partial H^*}{\partial t} + \frac{1}{2m}(\mathrm{grad}\ H^*)^2 + \frac{1}{2m}(\mathrm{grad}\ G^*)^2 - D\Delta G^* - Q = 0 \tag{4.1.7}$$

$$\frac{\partial G^*}{\partial t} + \frac{1}{m}\,\mathrm{grad}\ H^*\ \mathrm{grad}\ G^* - D\Delta H^* = 0 \tag{4.1.8}$$

Equations (4.1.7), (4.1.8) are equations (19), (20), of [4], except for the potential Q which was taken to be zero in [4].

At this stage, we make the following observation: if we define

$$\tilde{Q} = \frac{1}{m}(\nabla G^*)^2 - 2D\Delta G^* - Q \tag{4.1.9}$$

then (4.1.7) becomes

$$\frac{\partial H^*}{\partial t} + \frac{1}{2m} (\text{grad } H^*)^2 - \frac{1}{2m} (\text{grad } G^*)^2 + D\Delta G^* + \bar{Q} = 0 \tag{4.1.10}$$

(4.1.8) is unchanged:

$$\frac{\partial G^*}{\partial t} + \frac{1}{m} \text{grad } H^* \text{ grad } G^* - D\Delta H^* = 0 \tag{4.1.8}$$

If we now multiply (4.1.10) by i, and subtract (4.1.8), we obtain

$$\frac{\partial}{\partial t} (-G^*+iH^*) = - D\Delta H^* + \frac{1}{m} \text{grad } H^* \text{ grad } G^* + \frac{i}{2m} (\text{grad } G^*)^2$$

$$- \frac{i}{2m} (\text{grad } H^*)^2 - iD\Delta G^* - i\,\bar{Q}$$

or

$$\frac{\partial}{\partial t} (-G^*+iH^*) = iD\Delta(-G^*+iH^*) + \frac{i}{2m} (\text{grad}(-G^*+iH^*))^2 - i\tilde{Q} \tag{4.1.11}$$

Straightforward differentiation gives us

Proposition 4.1.A. If G, $\bar{G}$ are solutions to (4.1.1), (4.1.3), then

$$\psi = \exp\left(\frac{-G^*+iH^*}{\hbar}\right) \tag{4.1.12}$$

is a solution to

$$i\hbar \frac{\partial\psi}{\partial t} = \frac{-\hbar^2}{2m} \Delta\psi + \left(\frac{(\nabla G^*)^2}{m} - \frac{\hbar\Delta G^*}{m} - Q\right)\psi. \tag{4.1.13}$$

Conversely, suppose we start with the Schrödinger equation

$$i\hbar \frac{\partial\hat{\psi}}{\partial t} = \frac{-\hbar^2}{2m} \Delta\hat{\psi} - P\hat{\psi} \qquad (t,x) \in \bar{S} \tag{4.1.14}$$

with given potential P. Assume there is a solution $\hat{\psi}$ of (4.1.14), $\hat{\psi} \neq 0$, all (t,x), with

$$\hat{\psi} = \exp\left(\frac{-M+iN}{\hbar}\right) \tag{4.1.15}$$

and suppose that M and N are $C^{1,2}$ functions on $\bar{S}$. Running the above arguments backwards, we see

$$\frac{\partial N}{\partial t} + \frac{1}{2m} (\text{grad } N)^2 - \frac{1}{2m} (\text{grad } M)^2 + D\Delta M - P = 0 \tag{4.1.16}$$

$$\frac{\partial M}{\partial t} + \frac{1}{m} \text{grad } N \text{ grad } M - D\Delta N = 0 \tag{4.1.17}$$

The passage from (4.1.14) to the pair of equations (4.1.16), (4.1.17) was used by Louis de Broglie for introducing his "theorie du guidage" (see [5]; equations (4.1.16), (4.1.17) are the so-called equations (J) and (C) of Louis de Broglie). Together with this pair of equations he defined the quantum potential Q_p by

$$Q_p = D\Delta M - \frac{1}{2m}(\text{grad } M)^2 \qquad (4.1.18)$$

The purpose of the definition (4.1.18) was to reduce equation (4.1.16) to the form

$$\frac{\partial N}{\partial t} + \frac{1}{2m}(\text{grad } N)^2 + Q_p - P = 0 \qquad (4.1.19)$$

which is the Hamilton-Jacobi equation of classical mechanics for the motion of a mass-point in the potential $P - Q_p$. As the reader may anticipate, if we next introduce the "modified potential" $\hat{Q}$ by

$$\hat{Q} = P - 2Q_p = P - 2D\Delta M + \frac{(\text{grad } M)^2}{m} \quad , \qquad (4.1.20)$$

then

$$\frac{\partial}{\partial t}(N+M) + \frac{1}{2m}(\text{grad }(N+M))^2 - D\Delta(N+M) - \hat{Q} = 0 \qquad (4.1.21)$$

$$\frac{\partial}{\partial t}(M-N) - \frac{1}{2m}(\text{grad }(M-N))^2 + D\Delta(M-N) + \hat{Q} = 0 \qquad (4.1.22)$$

Thus we have

Proposition 4.1.B. If

$$\hat{\psi} = \exp\left(\frac{-M+iN}{\hbar}\right)$$

is a solution as above to

$$i\hbar\frac{\partial\psi}{\partial t} = \frac{-\hbar^2}{2m}\Delta\hat{\psi} - P\hat{\psi} \quad , \qquad (4.2.12)$$

then (M-N) is a solution of

$$\frac{\partial G}{\partial t} - \frac{1}{2m}(\text{grad } G)^2 + D\Delta G + \hat{Q} = 0 \quad , \qquad (4.1.23)$$

and (N+M) is a solution of

$$\frac{\partial\bar{G}}{\partial t} + \frac{1}{2m}(\text{grad } \bar{G})^2 - D\Delta\bar{G} - \hat{Q} = 0 \quad . \qquad (4.1.24)$$

Equations (4.1.23, (4.1.24) are the equations (4.1.1), (4.1.3) with Q replaced by $\hat{Q}$, which is given by (4.1.20) (note that $\hat{Q}$ is specified once P is given, and "subsequently" [illegible]

Remark 4.1.C

Nelson's Equations

If we take gradients of equations (4.1.8), (4.1.10) and define

$$u = \frac{-\nabla G^*}{m} , \quad \upsilon = \frac{\nabla V^*}{m}$$

we obtain

$$\frac{\partial \upsilon}{\partial t} = \frac{-\hbar}{2m} \Delta u + \frac{\text{grad } u^2}{2} - \frac{\text{grad } \upsilon^2}{2} - \frac{\nabla \tilde{Q}}{m} \qquad (4.1.25)$$

$$\frac{\partial u}{\partial t} = \frac{\hbar}{2m} \Delta \upsilon - \nabla(\upsilon \cdot u) \qquad (4.1.26)$$

Nelson derived these equations, which are the main dynamical equations in his theory theory, via different methods ($\tilde{Q}/m$ representing the force field acting on a microscopic microscopic particle undergoing a Brownian motion). He found, with

$$u = \frac{\hbar}{m} \nabla R$$

$$\upsilon = \frac{\hbar}{m} \nabla S$$

that

$$\psi = \exp(R+iS)$$

satisfied

$$i\hbar \frac{\partial \psi}{\partial t} = \frac{-\hbar^2}{2m} \Delta\psi + \tilde{Q}\psi$$

and the converse.

Doing this involved recognizing that (4.1.10), (4.1.8) (equivalently (4.1.17), (4.1.16) or C and J of Louis de Broglie) are the imaginary and real part of the Schrödinger equation (modulo the factor of $\bar{\psi}$) which we used in going from (4.1.10), (4.1.8) to Prop. 4.1.A.

4.2 Inhomogeneous "backward and forward" heat equations

Now, if we make the exponential transformation

$$\phi(t,x) = \exp(-G(t,x)/\hbar) \qquad (4.2.1)$$

in equation (4.1.1), we have

$$\frac{\partial \phi}{\partial t} = -D\Delta\phi + \frac{Q\phi}{\hbar} \qquad (4.2.2)$$

with

$$\phi(t_1,x) = \exp\left[-G\,\frac{(t_1,x)}{\hbar}\right] .$$

Similarly, if

$$\bar{\phi}(t,x) = \exp\left(\frac{-\bar{G}(t,x)}{\hbar}\right) \qquad (4.2.3)$$

is put in (4.1.3), we have

$$\frac{\partial\bar{\phi}}{\partial t} = -D\Delta\bar{\phi} + \frac{Q\bar{\phi}}{\hbar} \tag{4.2.4}$$

with

$$\bar{\phi}(t_0,x) = \exp\left[-\bar{G}\,\frac{(t_0,x)}{\hbar}\right]$$

Thus from Proposition (4.1.B) and the above transformation we have the following

Fact I. If $\hat{\psi}$ given by

$$\hat{\psi} = \exp\left[\frac{-M+iN}{\hbar}\right]$$

is a solution to (4.1.14) then

i) $\phi = \exp\left[\frac{-(M-N)}{\hbar}\right]$ is a solution of

$$\frac{\partial\phi}{\partial t} = -\ D\Delta\phi + \frac{\hat{Q}\phi}{\hbar} \tag{4.2.5}$$

ii) $\bar{\phi} = \exp\left[\frac{-(M-N)}{\hbar}\right]$ is a solution of

$$\frac{\partial\bar{\phi}}{\partial t} = +\ D\Delta\bar{\phi} - \frac{\hat{Q}\bar{\phi}}{\hbar} \tag{4.2.6}$$

iii) The square of the modulus of $\hat{\psi}(t,x)$ is given by

$$||\hat{\psi}(t,x)||^2 = \exp(-2M/\hbar) = \phi(t,x)\ \bar{\phi}(t,x) = \phi^*(t,x)$$

iv) ϕ^* is a solution of the Fokker-Planck equation

$$\frac{\partial\phi^*}{\partial t} = -\sum_{i=1}^{n}\frac{\partial}{\partial x_i}(v_i(t,x)\phi^*) + D\Delta\phi^* \tag{4.2.7}$$

where

$$v_i(t,x) = \frac{2D}{\phi(t,x)}\ \frac{\partial\phi(t,x)}{\partial x_i} \qquad i = 1,\ldots n. \qquad (t,x)\ \epsilon\ \bar{S}. \tag{4.2.8}$$

Conversely, suppose there exist solutions $\bar{\bar{\phi}}$, $\overset{\neg}{\phi}$ of the equations

$$\frac{\partial\tilde{\bar{\phi}}}{\partial t} = D\Delta\bar{\bar{\phi}} - \frac{R}{2mD}\,\bar{\bar{\phi}} \quad \text{in} \quad (t_0,t_1)\ x\ E^n \tag{4.2.9}$$

$$\frac{\partial\overset{\neg}{\phi}}{\partial t} = -D\Delta\overset{\neg}{\phi} + \frac{R}{2mD}\,\overset{\neg}{\phi} \quad \text{in} \quad (t_0,t_1)\ x\ E^n \tag{4.2.10}$$

for given R, satisfying conditions

$$\bar{\bar{\phi}}(t_0,\cdot) = \bar{\phi}_0 \tag{4.2.11}$$

$$\overset{\neg}{\phi}(t_1,\cdot) = \bar{\phi}_1 \tag{4.2.12}$$

where $\bar{\phi}_0$ and ϕ_1 are non-negative, continuous, a d bounded functions on E^n. (We refer to [8] for existence theory.) It may be seen, [8], that

$$\tilde{\bar{\phi}}(t,x) > 0, \quad \text{and } \tilde{\phi}(t,x) > 0 \text{ in } S$$

provided that neither $\bar{\phi}_0$ nor ϕ_1 vanishes identically. Now, defining $\tilde{\bar{W}}$, $\tilde{W}$ by

$$\tilde{\bar{\phi}} = \exp\left(-\frac{\tilde{\bar{W}}}{2mD}\right) \tag{4.2.13}$$

$$\tilde{\phi} = \exp\left(-\frac{\tilde{W}}{2mD}\right) \tag{4.2.14}$$

we see that $\tilde{W}$ is a solution of

$$\frac{\partial G}{\partial t} - \frac{1}{2m}(\text{grad } G)^2 + D\Delta G + R = 0 \tag{4.2.15}$$

with

$$G(t_1,x) = -\hbar \log \phi_1 \tag{4.2.16}$$

and $\tilde{\bar{W}}$ is a solution of

$$\frac{\partial \bar{G}}{\partial t} + \frac{1}{2m}(\text{grad } \bar{G})^2 - D\Delta G - R = 0 \tag{4.2.17}$$

with

$$\bar{G}(t_0,x) = -\hbar \log \bar{\phi}_1 \tag{4.2.16}$$

Thus, from Proposition 4.1.A and the above arguments, we have

<u>Fact II.</u> If $\tilde{\bar{\phi}}$, $\tilde{\phi}$ are solutions to the Cauchy problems (4.2.9), (4.2.11) and (4.2.10), (4.2.12), then

$$\tilde{\psi} = \exp\left(\frac{-(\tilde{\bar{W}}+\tilde{W}) + i(\tilde{\bar{W}}-\tilde{W})}{2\hbar}\right) = \exp\left(\frac{-\tilde{W}^* + i\tilde{V}^*}{\hbar}\right)$$

where

$$\tilde{W}^* = \frac{\tilde{\bar{W}}+\tilde{W}}{2}, \quad \tilde{V}^* = \frac{\tilde{\bar{W}}-\tilde{W}}{2},$$

satisfies

$$i\hbar \frac{\partial \tilde{\psi}}{\partial t} = \frac{-\hbar^2}{2m}\Delta\psi + \left(\frac{\left[\nabla\left((\tilde{\bar{W}}+\tilde{W})/2\right)\right]^2}{m} - \hbar\Delta\left(\frac{\tilde{\bar{W}}+\tilde{W}}{2m}\right) - R\right)\tilde{\psi} \tag{4.2.19}$$

Note:

a) the solution $\tilde{\psi}$ of the Schrödinger equation (4.2.19) depends, like $\bar{\bar{W}}$ and $\tilde{W}$, on the initial and terminal data of the Cauchy problems.

b)

$$||\tilde{\psi}(t,x)||^2 = \exp\left(\frac{-\bar{\bar{W}}(t,x) + \bar{W}(t,x)}{2mD}\right) = \bar{\bar{\phi}}(t,x)\,\tilde{\phi}(t,x) = \tilde{\phi}^*(t,x) \ . \tag{4.2.20}$$

Fact I is obtained in the proof of Theorem 4.3 of [15]; Fact II is more or less implicit in Theorem 4.4 and Corollary 4.4.1 of [15], however, the arguments here give Fact II more directly.

Example Homogeneous "backward and forward" heat equations, $n = 1$

The solution of

$$\frac{\partial\tilde{\phi}}{\partial t} = -D\Delta\tilde{\phi} \quad \text{on } [t_0,t_1] \text{ where } 0 < t_0 < t_1 < T, \tag{4.2.21}$$

$$\tilde{\phi}_1(x) = \frac{1}{\sqrt{4\pi D(T-t_1)}} \exp\left(\frac{-x^2}{4D(T-t_1)}\right) \tag{4.2.22}$$

is known to be

$$\tilde{\phi}(t,x) = \frac{1}{\sqrt{4\pi D(T-t)}} \exp\left(\frac{-x^2}{4D(T-t)}\right) \qquad t_0 \le t \le t_1 \tag{4.2.23}$$

Similarly, the solution of

$$\frac{\partial\bar{\bar{\phi}}}{\partial t} = D\,\frac{\partial^2\bar{\bar{\phi}}}{\partial t^2} \tag{4.2.24}$$

$$\bar{\bar{\phi}}_0(x) = \frac{1}{\sqrt{4\pi Dt_0}} \exp\left(\frac{-x^2}{4Dt_0}\right) \tag{4.2.25}$$

is known to be

$$\bar{\bar{\phi}}_0(t,x) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(\frac{-x^2}{4Dt}\right). \tag{4.2.26}$$

Then

$$\begin{aligned}\tilde{W}(t,x) &= -2mD \log \tilde{\phi}(t,x)\\ &= \frac{m}{2}\,\frac{x^2}{T-t} + mD\log(T-t) + mD\log 4\pi D\end{aligned} \tag{4.2.27}$$

$$\bar{\bar{W}}(t,x) = \frac{m}{2}\,\frac{x^2}{t} + mD\log t + mD\log 4\pi D \tag{4.2.28}$$

Now,

$$\tilde{W}^*(t,x) = \frac{(\bar{\bar{W}}+\tilde{W})}{2}(t,x) = \frac{mx^2}{4}\left(\frac{1}{T-t} + \frac{1}{t}\right) + \frac{mD}{2}\log(t(T-t)) + mD\log 4\pi D$$

$$\tilde{V}^*(t,x) = \frac{(\tilde{\bar{W}}-\tilde{W})}{2}(t,x) = \frac{mx^2}{4}\left(\frac{1}{t} - \frac{1}{T-t}\right) + \frac{mD}{2}\,(\log t - \log(T-t))$$

$$\nabla\tilde{W}^*(t,x) = \frac{mx}{2}\left(\frac{T}{t(T-t)}\right) \qquad \Delta\tilde{W}^*(t,x) = \frac{mT}{2t(T-t)}$$

$$\nabla\tilde{V}^*(t,x) = \frac{mx}{2}\left(\frac{T-2t}{t(T-t)}\right) \qquad \Delta\tilde{V}^*(t,x) = \frac{m(T-2t)}{2t(T-t)}$$

Thus, by Prop. 4.2.B,

$$\bar{\psi} = \exp\left\{ \frac{-\left(\frac{mx^2}{4}\left(\frac{1}{T-t} + \frac{1}{t}\right) + \frac{mD}{2}\log(t(T-t)) + mD\log 4\pi D\right)}{\hbar} + i\,\frac{\left(\frac{mx^2}{4}\left(\frac{T-2t}{t(T-t)}\right) + \frac{mD}{2}\log\left(\frac{t}{T-t}\right)\right)}{\hbar}\right\} \tag{4.2.29}$$

satisfies

$$i\hbar\,\frac{\partial\tilde{\psi}}{\partial t} = \frac{-\hbar^2}{2m}\Delta\tilde{\psi} + \left(\frac{mx^2}{4}\,\left(\frac{T}{t(T-t)}\right)^2 - \frac{\hbar T}{t(T-t)}\right)\tilde{\psi}. \tag{4.2.30}$$

References

1. A. Blaquiere, Liens entre la theorie geometrique des processus optimaux et la mecanique ondulatoire, C.R. Acad. Sc. Paris, Serie A., Vol. 262 (1966), pp. 539-595.

2. A. Blaquiere, Interpretation d'un coefficient de diffusion complexe en mecanique ondulatoire, C.R. Acad. Sc. Paris, Serie A, Vol 268 (1969), pp. 1304-1306.

3. A. Blaquiere, System Theory: A new approach to wave mechanics, J. Optim. Thy. Appl., 32, 4 (1980), pp. 463-478.

4. A. Blaquiere and A. Marzollo, An alternative approach to wave mechanics of a particle at the non-relativistic approximation, Information, Complexity and Control in Quantum Physics, Proc. of the 4th International Seminar on Mathematical Theory of Dynamical Systems and Microphysics, Udine, 1985, Springer-Verlag, Wien, 1987.

5. De Broglie, L., Une tentative d'interpretation causale et nonlineaire de la mecanique ondulatoire, Gauthier-Villars, Paris, 1956.

6. W. Fleming and R. Rishel, Deterministic and Stochastic Optimal Control, Springer-Verlag, Berlin, 1975.

7. F. Guerra and L. Morato, Quantization of dynamical systems and stochastic control theory, Physical Review D, 27, 8 (1983), pp. 1774-1786.

8. A.M. Il'in, A.S. Kalashnikov, O.A. Oleinik, Linear Equations of a Second Order of Parabolic Type, Russian Mathematical Surveys, Vol. 17, Macmillan and Co., Ltd., London, 1962.

9. S. Mitter, Non-linear Filtering and Stochastic Mechanics, Stochastic Systems: The Mathematics of Filtering and Identification with Applications, Proc. NATO Advanced Study Institute, Les Arcs, Savoie, France 1980, Reidel, Dordrecht, 1981.

10. E. Nelson, Derivation of the Schrödinger Equation from Newtonian Mechanics, Physical Review, 150, 4 (1966), pp. 1079-1085.

11. E. Nelson, Quantum Fluctuations, Princeton U.P., Princeton, 1985

12. L. Papiez, Stochastic optimal control and quantum mechanics, J. Math. Phys., 23, 6 (1982), pp. 1017-1019.

13. K. Yasue, Quantum mechanics and stochastic control theory, J. Math. Phys., 22, 5 (1981), pp. 1010-1020.

14. K. Yasue, Stochastic Calculus of Variations, J. Func. Analysis, 41 (1981), pp. 327-340.

15. J.C. Zambrini, Variational processes and stochastic versions of mechanics, J. Math. Phys., 27, 9 (1986), pp. 2307-2330.

CONTINUOUS PROGRAMMING AND NONLINEAR FILTERING OF QUANTUM CONTROLLED PROCESSES.

V.P. BELAVKIN
Moscow Institute of Electronic Mashinebuilding
B. Vusovski 3/12, Moscow 109028

A quantum continuous Bellman equation is derived for the solution of the problem of optimal control of a quantum stochastic process with nondemolition measurements. The solution of this equation $u^o(t,u^t,\rho)$ together with the solution of the corresponding nonlinear filtering problem $\rho = \hat{\pi}(t)$ defines the optimal control strategy $d^o(t,z^t,q(t)) = u^o(t,u^t,\hat{\pi}(t))$.

Let us consider a quantum controlled process over the algebra $\mathcal{E} = \mathcal{B}(E)$ described by the family of normal representations $i(t) : \mathcal{E} \mapsto \mathcal{B}_t \otimes C(\mathcal{U}^t)$ where $\mathcal{B}_t = \mathcal{E} \otimes \mathcal{B}(F^t)$, $F^t = \Lambda\ (\mathcal{L}^2([0,t[)$ is the Fock space.

Let $\mathcal{U}^t \subseteq \underset{t'<t}{\times} \mathcal{U}(t')$ be a Hausdorf space of controlling processes $u^t = \{u(t') | t' \in [0,t[\}$ such that $\mathcal{U}^t \times \mathcal{U}^s_t = \mathcal{U}^{t+s}$ for all $t,s \in R_+$, where $\mathcal{U}^s_t \subseteq \underset{0 \leq \tau < s}{\times} \mathcal{U}(t+\tau)$ and $\mathcal{U}_t = \mathcal{U}^\infty_t$, $\mathcal{U} = \mathcal{U}^\infty_o$. We consider a quantum controlled process $i(t,u^t) = i(t)(u^t)$ over the algebra $\mathcal{E} = \mathcal{B}(E)$ with respect to $\mathcal{B}_t = \mathcal{E} \otimes \mathcal{B}(F^t)$, F^t is the Fock space over $\mathcal{L}^2([0,t[)$ described by the Hudson-Parthasarathy dynamical equation (1) for $P(t,u^t) = i(t,u^t,\rho)$

$$dP - \gamma(u,P) \otimes Idt = 2\mathrm{Re}\beta(P) \otimes dA + \Delta(P) \otimes dN, \qquad P(0) = p \in \mathcal{E} ,$$

where $\Delta = \Delta(t,u^t)$, $\beta = \beta(t,u^t)$, $\gamma(u(t)) = \gamma(t,u^t,u(t))$ are defined in standard way by operator-valued functions $V(t)$, $X(t) : \mathcal{U}^t \to A(t,u^t) = i(t,u^t,\mathcal{E})$ with unitary $V(t,u^t)$ and self-adjoint $H(t,u^t,u(t))$. We shall suppose that the controls $u_t \in \mathcal{U}_t$ are defined by strategies $u_t = d_t(z^t,q_t) = \{ d_t(t+\tau) | \tau \in [0,s[\}$ where $z^t = (u^t,q^t)$, $q_t = q^\infty_t$, $q^t = q^t_o$, $q^s_t = \{q(t+\tau) | \tau \in [0,s[\}$ are the results of nondemolition measurements on the interval $[t,t+s[$, $q^s_t \in R^{[t,t+s[}$ described by a commutative process $Q(t)$ satisfying the equation respecting to $dY = X \otimes Idt + V \otimes dA$, $d\Pi = X^* X \otimes Idt + 2\mathrm{Re}X^* V \otimes dA + I \otimes dN$

$$dQ - g(u(t)) \otimes Idt = 2\mathrm{Re}(b \otimes I)dY + (f \otimes I)d\Pi , \qquad Q(0) = x 1$$

with $b(t,u^t)$, $f(t,u^t)$, $g(t,u^t,u(t)) \in A(t,u^t)'$.

Let us consider the optimal control problem with the operator-valued risk $R_t(u) \in A_t(u) = \bigvee_{s>0} A(t+s,u^{t+s})$, satisfying the equation

$$R_{t_o}(u) = \int_{t_o}^{t_1} S(t,u^t,u(t))dt + R_{t_1}(u) ,$$

where $S(t,u^t,u(t)) \in A(t,u^t)$ for all $t \in R_+$. The optimal control strategy d_t^o is defined as a solution of the extremal problem

$$\langle \rho \otimes \omega, R_t(u^t,d_t(z^t,q_t)) \rangle = \inf ,$$

where ρ is an initial state on $\mathcal{A}$ and ω is the vacuum state on $\mathcal{B}(F), F = F^\infty$. This solution can be found by the quantum dynamic programming method as the solution of the following Bellman continuous inverse-time equation.

THEOREM. Let $r(t,z^t,d_t) \in \mathcal{A}$ be the averaged risk operator uniquely defined by

$$i(t,u^t,r(t,z^t,d_t)) = \mathcal{E}_t[R_t(u^t,d_t(z^t,q_t))]$$

where $\mathcal{E}_t$ is the conditional expectation with respect to $\mathcal{B}_t = \mathcal{A} \otimes \mathcal{B}(F^t)$ corresponding to the vacuum state ω_t on $\mathcal{B}(F_t)$ and

$$\hat{r}(t,z^t,d_t) = \mathcal{X}_t \circ i(t,u^t,r(t,z^t,d_t)) = \langle \hat{\pi}(t,z^t),r(t,z^t,d_t) \rangle$$

be the posterior risk, corresponding to the strategy d_t, where $\mathcal{X}_t$ is the conditional expectation on $\mathcal{B}_t$ with respect to the commutative algebra $\mathcal{C}_t$ generated by $q^t = \{q(t') | t' < t\}$. Then $\inf\limits_{d_t} \langle \hat{\pi}(t,z^t),r(t,d_t) \rangle = r(t,u^t,\hat{\pi}(t,z^t))$ where the functional $\rho \mapsto r(t,u^t,\rho)$ satisfies the following Bellman equation :

$$-\partial_t r(\rho) = \inf_{u \in \mathcal{U}(t)} \{\langle \rho,s(u) \rangle + (\langle \rho \circ \gamma(u),\delta \rangle + \frac{1}{2}(|b|^2 + f^2 \langle \rho \circ \alpha,\delta \rangle^2)r(\rho)\}$$

where $\partial_t = \partial/\partial t$, $\delta = \delta/\delta\rho$, $x \in \mathcal{A} : i(t,u^t,x) = X(t,u^t)$,

$$\rho \circ \gamma(u) = i[\rho,h(u)] + \frac{1}{2}([x\rho,x^*] + [x,\rho x^*]) ,$$

$$\rho \circ \alpha = 2\mathrm{Re}\, b(x - \langle \rho,x \rangle)\rho + f(x\rho x^* - \langle \rho,x^*x \rangle \rho),$$

and $s(u)$, $h(u) \in \mathcal{A}$ are defined by

$$i(t,u^t,s(u)) = S(t,u^t,u),$$

$$i(t,u^t,h(u)) = H(t,u^t,u)$$

and $\hat{\pi}(t,z^t)$ is an posterior state on $\mathcal{A}$ satisfying the nonlinear filtering equation

$$d\hat{\pi} - \hat{\pi} \circ \gamma(u)dt = \hat{\pi} \circ \alpha d\tilde{Q}/(|b|^2 + f^2 \langle \hat{\pi},x^*x \rangle) , \hat{\pi}(0) = \rho,$$

where $d\tilde{Q} = 2\mathrm{Re}\, b d\tilde{Y} + f d\tilde{\Pi}$, $d\tilde{Y} = dY - \langle \hat{\pi},x \rangle dt$, $d\tilde{\Pi} = d\Pi - \langle \hat{\pi},x \rangle dt$.

In particular, for the Brownian observation ($f = 0$)

$$-\partial_t r(\rho) = \inf_u \{\langle \rho,s(u) \rangle + (\langle \rho \circ \gamma(u),\delta \rangle + 2\langle \mathrm{Re}\,\theta(x - \langle \rho,x \rangle)\rho,\delta \rangle^2)r(\rho) \}$$

where $\theta = b/|b|$ and for the Poissonian observation $(b = 0)$

$$-\partial_t r(\rho) = \inf_u \{ \langle \rho, s(u) \rangle + \langle\langle \rho \circ \Upsilon(u), \delta \rangle + \frac{1}{2} \langle \rho - x\rho x^* / \langle \rho, x^* x \rangle, \delta \rangle^2) r(\rho) \} \ .$$

The linear dynamical programming for Gaussian ρ and canonical x was considered in (2), and the general formulation of quantum dynamical programming for the partially observable controlled quantum objects in operational approach was given in (3) .

REFERENCES

1. R.L. HUDSON, K.R. PARTHASARATHY, Quantum Ito's formula and stochastic evolutions, Comm. Math. Phys., 93, 1984, p.301-323 .
2. V.P. BELAVKIN, Nondemolition measurement and control in quantum dynamical systems. Proc. of 4th Int. Seminar in Math. Theory of dynamical systems and Microphysics : "Information complexity and control in quantum physics" , Udine 1985, Edts A. Blaquière, S. Diner, G. Lochak, Springer Verlag, Wien - New York, 1987, p.311-336.
3. V.P. BELAVKIN, Theory of the control of observable quantum qyqtems, Automatica and Remove Control, 44 (2), 1983, p.178-188.

MODELISATION ET COMMANDE DES SYSTEMES BIOLOGIQUES ET DES ECOSYSTEMES

MODELS AND CONTROL POLICIES FOR BIOLOGICAL SYSTEMS AND ECOSYSTEMS

AUTOMATIQUE ET REGULATION BIOLOGIQUE

Daniel CLAUDE

Laboratoire des Signaux et Systèmes,
C.N.R.S.- E.S.E.,
Plateau du Moulon, 91190 Gif-sur-Yvette, France.

Résumé : A la mémoire de Richard Bellman, nous présentons les contrôles bipolaires en biologie. De par ses seules applications thérapeutiques aux domaines des tumeurs cérébrales et de la cancérologie, cette méthodologie, liant l'automatique à la régulation biologique, aurait certainement eu ses faveurs. Nous en montrons toute la richesse en ouvrant d'autres perspectives qui justifient pleinement le lien entre les mathématiques et la médecine qui intéressait tant Richard Bellman.

Abstract : In memory of Richard Bellman, we present bipolar controls in biology. From its therapeutic applications in the field of cerebral tumors and cancerology alone, Richard Bellman would have certainly been in favour of this methodology which links control theory to biological regulation. We show all its richness in opening other prospects that entirely justify the link between mathematics and medicine which interested him so much.

I . INTRODUCTION

Depuis maintenant plusieurs décennies, de nombreux chercheurs ont pensé à créer un lien entre les mathématiques et la médecine (cf. les livres récents de Winfree [26] et de Swan [25]), en particulier par les essais de modélisation de certains phénomènes biologiques et par exemple, en cancérologie , par la recherche de procédures médicamenteuses (chimiothérapie) ou par la mise en place de protocoles d'émission de particules actives spécifiques (radiothérapie). Ils souhaitaient ainsi réunir la théorie mathématique et la pratique médicale. L'automatique, appliquée à certaines régulations biologiques, répond à cette exigence et à cette espérance.

En biologie, de nombreuses régulations font appel à plusieurs agents aux actions couplées. Il en est ainsi de la régulation de l'hydratation cellulaire ou du contrôle de la mitose dans lesquels interviennent les corticoïdes d'une part et la vasopressine d'autre part, de même que l'insuline et le glucagon régulent l'activité glycémique. La faillite dans certaines pathologies des thérapeutiques consistant à administrer une seule hormone trouve son explication dans le fait que l'on a négligé les réactions de l'autre hormone qui intervient à cause d'un jeu subtil de feedbacks croisés. En outre, la biologie est un domaine fortement non linéaire où le principe de superposition des actions n'a pas cours.

Ainsi, toute action thérapeutique mesurable doit passer par une modélisation non linéaire multivariable, suffisamment riche pour prendre en compte les aspects prépondérants des phénomènes étudiés, et assez simple pour envisager d'une manière raisonnable les possibilités de commande de ces systèmes et en déduire les actions thérapeutiques. A cause des couplages, les solutions proposées, par leur caractère faussement paradoxal, peuvent surprendre, déranger, voir provoquer des hostilités. Pourtant, les résultats cliniques sont là, authentifiés par les radiographies et les scanners, et on doit espérer que les deux exemples que nous allons traiter, permettent de convaincre de la nécessité de développer rapidement le champ d'action des thérapeutiques bipolaires dont Bernard-Weil est à l'origine.

II LE SYSTEME SURRENO-POSTHYPOPHYSAIRE ET LA VASOPRESSINO-CORTICOTHERAPIE

Dans le cadre de l'application de l'automatique aux traitements chimiothérapiques en cancérologie, Sundareshan et Fundakowski [24], s'interrogent sur le caractère dual de l'objet de ces thérapeutiques et souhaitent trouver des agents qui soient capables de détruire les cellules malignes tout en préservant les cellules saines. En fait, au sein de l'organisme existe un important système qui assure la régulation du développement cellulaire tant au point de vue de la mitose que de l'hydratation de la cellule, c'est le système hormonal surréno-posthypophysaire.

Le système surréno-posthypophysaire, formé par les cortico-surrénales d'une part et par la neuro-posthypophyse d'autre part, intervient ainsi au premier chef dans les manifestations cliniques observées chez le malade neuro-chirurgical. Ce système est responsable de manifestations aussi diverses que certains oedèmes du cerveau, certains collapsus cérébraux aggravant les suites d'intervention pour hématome sous-dural, et intervient dans l'évolution des tumeurs cérébrales malignes.

La reconnaissance du couplage entre ces deux glandes date des années 30 (cf.[23]), et ce système, aux actions ago-antagonistes (cf.[4, 6, 7]), assure des régulations majeures. Ainsi, la cortisone, secrétée par les cortico-surrénales, est un merveilleux agent, non seulement contre l'hyperhydratation cellulaire mais aussi comme produit anti-mitotique, comme cela a été démontré *in vitro* aussi bien dans le cas de tumeurs cérébrales malignes en culture que dans celui de toute autre lignée cancéreuse en culture de tissu. Quant à la vasopressine, secrétée par la neuro-posthypophyse, elle est responsable de la réabsorption de l'eau par le tube rénal et est un facteur de croissance tout à fait important. Ce premier facteur de croissance polypetidique a été découvert en 1968 par Bernard-Weil, Dalage, Olivier et Piette [9] et leur résultat a été confirmé ultérieurement par les auteurs américains, Rozengurt et all. [20], en 1979, et Monaco et all. [18] en 1982. Nous renvoyons à Pawlikowski [19] pour avoir un rappel récent des actions mitogéniques des neuropeptides. Le déséquilibre entre les corticoïdes et la vasopressine, avec un excès de vasopressine favorisant le développement tumoral, a été de nouveau mesuré récemment en cancérologie digestive (cf. [11]), mais il a été constaté dans bien d'autres cas. De plus, à cause du couplage entre ces deux hormones, certains oedèmes cérébraux résistent à la cortisone et les tumeurs cancéreuses ne sont vraiment influencées par les corticoïdes que pour un court laps de temps et avec des doses très élevées de ces hormones. Pire encore, l'organisme malade se place dans une position " d'homéostasie pathologique" (cf. Bernard-Weil [4]) et ce déséquilibre régulé bénéficie de sauvegardes biologiques puissantes qui tendent à le maintenir en l'état comme s'il s'agissait du fonctionnement physiologique " normal".

C'est ainsi qu'est préservé le déséquilibre vasopressine-corticoïdes chez le malade cancéreux, l'administration de corticoïdes ayant pour effet d'augmenter le taux de vasopressine pourtant déjà anormalement élevé (cf. [3]).

La solution consiste donc à envisager l'administration simultanée de vasopressine et de corticoïdes (cf.[5]), un modèle multivariable non linéaire venant conforter les intuitions premières du médecin (cf.[4, 6, 7]).

Ce modèle, représenté par un système différentiel non linéaire à deux entrées, e_1 et e_2, et deux sorties, z_1 et z_2, peut s'écrire sous la forme suivante (cf. [15]) :

$$\begin{aligned}
\dot{\mathfrak{X}} &= \sum_{i=1}^{3} [\, k_i (u+p)^i + c_i (v+q)^i \,] + e_1 \\
\dot{\mathfrak{Y}} &= \sum_{i=1}^{3} [\, k_i' (u+p)^i + c_i' (v+q)^i \,] + e_2 \\
\dot{X} &= e_1 \\
\dot{Y} &= e_2 \\
z_1 &= \mathfrak{X} - \mathfrak{Y} \\
z_2 &= m \operatorname{Log}[(\mathfrak{X}+\mathfrak{Y})/m]
\end{aligned} \tag{2.1}$$

avec $\mathfrak{X} = x+X$; $\mathfrak{Y} = y+Y$, où x et y désignent respectivement les actions des corticoïdes et de la vasopressine endogènes et X et Y les actions des hormones exogènes (thérapeutique).

Il s'agit d'un développement en série dans lequel apparaissent des expressions antagonistes $u = \mathfrak{X}-\mathfrak{Y}$ et des expressions agonistes $v = m \operatorname{Log} [(\mathfrak{X}+\mathfrak{Y})/m] + \theta(t)$, avec $\theta(t) = A + B \sin(\omega t) + C \cos(\omega t)$, où les constantes A, B, C et ω ($\omega = 2\pi / 24$ dans un rythme circadien) déterminent le synchroniseur $\theta(t)$ lié aux rythmes biologiques. L'introduction de la puissance cubique se justife par les conditions de stabilité du système (cf. [4]) ; p(t) représente un possible stimulus osmotique ; q(t) correspond à un éventuel stimulus volémique (hémorragie par exemple) ou un stress ; les paramètres, k_i, c_i, k_i', c_i' (i =1,2,3) sont constants ; le paramètre m est pris en général constant (m = 0,8) mais peut aussi être considéré comme variable dans le temps. Lorsque q a des valeurs positives, par forte augmentation de la volémie par exemple, et telles que x et y deviendraient négatifs, on prévoit la possibilité de faire quitter à m la valeur 0,8 pendant le transitoire nécessaire.

Le système est écrit dans un système d'unité commune (u.c.) pour lequel :

0,4 u.c. = 77 ng/ml de cortisol (F) = 1,1 µU/ml de vasopressine (VP),

valeurs qui correspondent à la moyenne des valeurs expérimentales des rythmes circadiens de ces hormones.

Les valeurs x, y, X, Y peuvent être assimilées à des concentrations hormonales et sont ainsi sujettes à des contraintes de positivité. Dans le cas physiologique ($X = 0$, $Y = 0$; $p = 0$, $q = 0$), l'équilibration est simulée avec un champ paramétrique de (2.1) donnant un cycle-limite tel que le couple (u,v) admette l'origine (0, 0) comme point critique. L'équilibration ($X = 0$, $Y = 0$) devient pathologique si une modification du champ (2.1) permet à un nouveau cycle-limite d'apparaître.

Les paramètres k_i, c_i, k_i', c_i' ($i = 1, 2, 3$) pour le système simulant la pathologie, et $\bar{k}_i$, $\bar{c}_i$, $\bar{k}_i'$, $\bar{c}_i'$ pour le système simulant le cycle physiologique, sont identifiés à partir des données cliniques et physiologiques, à l'aide de la méthode d'intégration numérique de Davidon-Fletcher-Powell avec contraintes (cf. [1]). Le critère à minimiser $\mathfrak{J}$ (k_i, c_i, k_i', c_i', T) est donné par :

$$\mathfrak{J}\left(k_i, c_i, k_i', c_i', T \right) = \sum_j \left[(\bar{x}_j - x_j)^2 + (\bar{y}_j - y_j)^2 \right] \qquad (2.2)$$

où $\bar{x}$ et $\bar{y}$ désignent des valeurs expérimentales et x et y les solutions "endogènes" du système (2.1) dans lequel on a pris $X = 0, Y = 0$, $p = 0$ et $q = 0$. La quantité T correspond à trois cycles, soit ici, à 72 heures.

Dans le cas d'une homéostatie pathologique, la "simulation thérapeutique" consiste à déterminer les hormones exogènes X et Y de façon à ramener le système dans une position d'homéostasie physiologique.

Une première méthode (cf. [6, 7]) consiste à écrire les entrées e_1 et e_2 dans une forme semblable à celle des hormones endogènes, soit :

$$\begin{aligned} e_1 &= \sum_{i=1}^{3} \left[k_{3+i} \, (u + p)^i + c_{3+i} \, (v + q)^i \right] + \lambda_1 (X - \alpha_1) + \lambda_2 (X - \alpha_1)^2 + \lambda_3 (X - \alpha_1)^3 \\ e_2 &= \sum_{i=1}^{3} \left[k_{3+i}' \, (u + p)^i + c_{3+i}' \, (v + q)^i \right] + \lambda_1' (Y - \alpha_1') + \lambda_2' (Y - \alpha_1')^2 + \lambda_3' (Y - \alpha_1')^3 \end{aligned} \qquad (2.3)$$

avec λ_1, λ_2, λ_3, λ_1', λ_2', λ_3', α_1, α_1' des paramètres constants ayant pour rôle d'éviter la dérive du cycle-limite de dimension 4 que suivent les quatre états du système. On identifie alors les paramètres de (2.3) à l'aide de la méthode de Davidon-Fletcher-Powell.

Remarque 1: La tentation de prendre pour les entrées e_1 et e_2 la différence entre les équations de l'état physiologique et de l'état pathologique conduit à un contrôle qui peut ne pas satisfaire les conditions de positivité des variables x, y, X, Y, ni assurer l'existence d'un cycle-limite (cf. [6]).

Une seconde méthode, basée en premier (cf. [15]) sur le découplage et la linéarisation des systèmes non linéaires (cf. [12, 13, 14] et les bibliographies afférentes), consiste en fait à inverser le système (2.1) (cf. [16, 17]).

A partir du système (2.1), on considère alors les relations suivantes :

$$H = 1/2 \left(z_1 + m\, e^{(z_2/m)} \right)$$

$$\Psi = 1/2 \left(- z_1 + m\, e^{(z_2/m)} \right)$$

$$X = H - x \qquad (2.4)$$

$$Y = \Psi - y$$

x et y étant solutions des équations différentielles :

$$\dot{x} = \sum_{i=1}^{3} [\, k_i (z_1 + p)^i + c_i (z_2 + \theta + q)^i \,]$$

$$\dot{y} = \sum_{i=1}^{3} [\, k_i^{'} (z_1 + p)^i + c_i^{'} (z_2 + \theta + q)^i \,] \qquad (2.5)$$

Il s'agit donc de permettre aux sorties z_1 et z_2 du système (2.1) de passer de la position pathologique, donnée par les équations différentielles :

$$\dot{\psi}_1 = \sum_{i=1}^{3} [\, (k_i - k_i^{'}) (\psi_1 + p)^i + (c_i - c_i^{'}) (\cdot\cdot + q)^i \,]$$

$$\dot{\psi}_2 = \left(\sum_{i=1}^{3} [\, (k_i + k_i^{'}) (\psi_1 + p)^i + (c_i + c_i^{'}) (v + q)^i \,] \right) . e^{(-\Psi_2/m)} \qquad (2.6)$$

avec $v = \psi_2 + A + B \sin(\omega t) + C \cos(\omega t)$ et $\omega = 2\pi / 24$,

à l'équilibre physiologique décrit par les équations différentielles obtenues à partir des données expérimentales :

$$\dot{\varphi}_1 = \sum_{i=1}^{3} [\, (\bar{k}_i - \bar{k}_i^{'}) \varphi_1^i + (\bar{c}_i - \bar{c}_i^{'}) v^i \,]$$

$$\dot{\varphi}_2 = \left(\sum_{i=1}^{3} [\, (\bar{k}_i + \bar{k}_i^{'}) \varphi_1^i + (\bar{c}_i + \bar{c}_i^{'}) v^i \,] \right) . e^{(-\varphi_2/m)} \qquad (2.7)$$

avec $v = \varphi_2 + A + B \sin(\omega t) + C \cos(\omega t)$ et $\omega = 2\pi / 24$.

On écrit z_1 et z_2 sous la forme :

$$\begin{aligned} z_1 &= \delta_1 + \varphi_1 \\ z_2 &= \delta_2 + \varphi_2 \end{aligned} \qquad (2.8)$$

Le souhait du thérapeute est alors de trouver des fonctions δ_1 et δ_2 qui permettent en premier de définir un transitoire amenant les courbes pathologiques initiales, représentées par x et y, vers les courbes physiologiques que doivent suivre les variables $\mathcal{X}$ et $\mathcal{Y}$, somme des actions des hormones endogènes et exogènes. En second, après la période transitoire (deux à trois jours), le thérapeute souhaite à la fois, voir s'installer un régime permanent aussi proche que possible du rythme circadien physiologique pour les variables $\mathcal{X}$ et $\mathcal{Y}$, et mettre en place, pour de nombreuses raisons faciles à deviner, une action thérapeutique périodique de période égale ici à 24 heures.

Cependant, l'analyse immédiate des équations (2.5) montre qu'avec les coefficients du pathologique, il n'y a aucune raison pour que l'introduction dans ces équations des rythmes physiologiques entraîne l'apparition d'un cycle-limite. Bien au contraire, comme le confirment les simulations numériques, on assiste à une dérive affine du cycle. La démonstration de ce phénomène étant évidente.

Ainsi, la seule possibilité, en régime permanent, est de déformer aussi peu que possible le rythme physiologique pour assurer la périodicité de la thérapeutique représentée par X et Y, les fonctions δ_1 et δ_2 étant alors elles aussi périodiques. On est conduit ainsi à réaliser une optimisation sous les contraintes $x \geq 0$, $y \geq 0$, $X \geq 0$, $Y \geq 0$. Enfin, il faut s'assurer que le cycle-limite obtenu est stable et que le système est en plus structurellement stable.

Il est à noter que le principe de l'utilisation de l'optimisation est judicieux au regard de la notion de rythme physiologique moyen qui est utilisée et aussi vis à vis des incertitudes qu'amène l'utilisation d'un modèle. Ainsi les fonctions δ_1 et δ_2 doivent permettre de satisfaire les conditions de positivité des variables x, y, X, Y, et, après un transitoire, doivent assurer l'existence d'un régime permanent cyclique et basé sur le rythme circadien. Il s'agit alors de trouver une classe de fonctions suffisamment riche pour pouvoir contenir les solutions cherchées. On peut envisager une recherche hybride en séparant la partie transitoire du régime permanent.On peut alors considérer la classe de fonctions à quatre paramètres, dense dans l'ensemble des fonctions continues sur tout intervalle compact, et définie par :

$$f(x) = \int_0^{x+a} \frac{bd}{1 + d^2 - \cos(bt)} \cos(e^t)\,dt + c \quad \text{avec } d > 0 \qquad (2.9)$$

Cette classe de fonctions est utilisée par Boshernitzan [2] dans la recherche des équations différentielles universelles (cf. [2, 22]).

Cette seconde méthode est en cours d'étude.

Remarque 2 : On pourrait s'inquiéter de l'impossibilité de trouver un contrôle thérapeutique capable de rétablir les rythmes physiologiques, mais on ne doit pas oublier que dans la réalité les paramètres qui déterminent le comportement du système sont variables et si ils sont passés de la position physiologique à la situation pathologique, le thérapeute, dans les cas de réversibilité, postule qu'un maintien forcé d'un rythme proche du rythme physiologique, pendant une période suffisante, permettra aux paramètres de se recaler sur l'homéostasie physiologique.

III. LE COUPLAGE INSULINE-GLUCAGON ET LE DIABETE.

L'activité glycémique peut être considérée comme la résultante des actions antagonistes du glucagon hyperglycémiant et de l'insuline hypoglycémiante, ces deux hormones agissant d'une façon couplée. Ce système présente, par rapport au système surréno-posthypophysaire, une particularité remarquable au plan anatomique. Dans le cas de la réponse glycémique, la nature a installé le mécanisme de commande de la régulation dans un même endroit - les îlots de Langerhans - au sein du pancréas. On trouve dans ces amas cellulaires la fabrication simultanée de l'insuline et du glucagon sous l'action coordonnée de la somatostatine. Devant les résultats cliniques obtenus à l'aide de la vasopressino-corticothérapie, il semblait, au regard des enjeux en diabétologie, interessant de proposer une modélisation du système insuline-glucagon sous l'angle de la vision bipolaire des systèmes ago-antagonistes définis par Bernard-Weil.

La modélisation proposée prend la forme d'un système différentiel non linéaire, à trois entrées e_1, e_2, e_3 et trois sorties z_1, z_2, z_3 ainsi défini (cf. [8]) :

$$\dot{\mathfrak{X}} = \sum_{i=1}^{3} [k_i (\mathfrak{X} - \mathfrak{Y} + e_3)^i + c_i (\mathfrak{X} + \mathfrak{Y} - m)^i] + e_1$$

$$\dot{\mathfrak{Y}} = \sum_{i=1}^{3} [k_i' (\mathfrak{X} - \mathfrak{Y} + e_3)^i + c_i' (\mathfrak{X} + \mathfrak{Y} - m)^i] + e_2$$

$$\dot{X} = e_1$$

$$\dot{Y} = e_2 \tag{3.1}$$

$$\dot{G} = g_1 (G_0 - G) + g_2 [g_3 [\mathrm{th}(g_4 (\mathfrak{X} - \mathfrak{Y} + Y - X + g_5)) + \mathrm{th}(g_4 (X - Y + g_5)) - 2\mathrm{th}(g_4 g_5)] + e_3]$$

$$z_1 = \mathfrak{X} - \mathfrak{Y}$$

$$z_2 = \mathfrak{X} + \mathfrak{Y} - m$$

$$z_3 = G$$

avec $\mathcal{X} = x + X$, $\mathcal{Y} = y+Y$, où x et y désignent respectivement les actions du glucagon et de l'insuline endogènes et X et Y les actions des hormones exogènes (thérapeutique).
$G_0 = 0{,}78$, désigne le taux de base physiologique de la réponse glycémique $G(t)$ et $m = 2{,}1$.

Le système est écrit dans un système d'unité commune pour lequel une unité commune vérifie :
10 µU/ml d'insuline = 100 pg/ml de glucagon.
Dans le cas de l'étude du test de tolérance au glucose, l'entrée $e_3 = p(t)$, qui est liée à la prise orale de 100 g de glucose, est représentée par la fonction :

$$p(t) = (p_1 / (p_1 - p_2)) \,.\, 100 \,.\, p_3 \,[\, \exp(-p_2 t) - \exp(-p_1 t) \,] \qquad (3.2)$$

Comme pour le modèle du système surréno-pothypophysaire, les paramètres des équations (3.1) et (3.2) ont été identifiés, à l'aide de la méthode d'optimisation non linéaire de Davidon-Fletcher-Powell, à partir des courbes expérimentales. Les paramètres définissant la fonction $p(t)$ ont été ajustés une seule fois car les conditions d'absorption intestinale du glucose sont moins influencées par les anomalies hormonales que les autres processus du métabolisme glucidique. Par contre, bien entendu, les paramètres de l'équation (3.1) sont à identifier dans le cas physiologique et dans le cas pathologique.
La recherche du contrôle (thérapeutique) visant à corriger les anomalies de la réponse glycémique chez le diabétique a été obtenue dans un premier temps (cf. [8]) en prenant les entrées e_1 et e_2 sous la forme :

$$\begin{aligned} e_1 &= \sum_{i=1}^{3} [\, k_{3+i} (\mathcal{X} - \mathcal{Y} + p)^i + c_{3+i} (\mathcal{X} + \mathcal{Y} - m)^i \,] \\ e_2 &= \sum_{i=1}^{3} [\, k'_{3+i} (\mathcal{X} - \mathcal{Y} + p)^i + c'_{3+i} (\mathcal{X} + \mathcal{Y} - m)^i \,] \end{aligned} \qquad (3.3)$$

Elles permettent de mettre en place un contrôle asymptotique tendant à ramener la position limite pathologique à la valeur physiologique moyenne de la glycémie (1 g/l), le déséquilibre initial glucagon-insuline avant la charge en glucose, comme l'équilibre physiologique, étant des points critiques stables du modèle physiologique.
On peut aussi opérer comme pour le système surréno-posthypophysaire et considérer les relations :

$$\begin{aligned} \mathcal{X} &= 1/2\,(z_1 + z_2 + m) \\ \mathcal{Y} &= 1/2\,(-z_1 + z_2 + m) \\ X &= \mathcal{X} - x \\ Y &= \mathcal{Y} - y \end{aligned} \qquad (3.4)$$

x et y étant solutions des équations différentielles :

$$\dot{x} = \sum_{i=1}^{3} [\, k_i (z_1 + p)^i + c_i (z_2\,)^i \,]$$

$$\dot{y} = \sum_{i=1}^{3} [\, k_i' (z_1 + p)^i + c_i' (z_2)^i \,] \qquad (3.5)$$

Il s'agit ici de permettre aux sorties z_1, z_2 et z_3 du système (3.1) de passer de la position pathologique :

$$z_1(0)\,;\, z_2(0)\,;\, G(0) \qquad (3.6)$$

à l'équilibre physiologique " asymptotique" :

$$z_1 = 0\,;\, z_2 = 0\,;\, G = 1 \qquad (3.7)$$

L'équilibre physiologique devant bien entendu être atteint avant l'ingestion suivante, soit dans un délai d'environ 5 heures.

Pour déterminer la "thérapeutique" - X, Y - à appliquer au système "pathologique" (3.1) on peut alors, par exemple, utiliser de nouveau la classe de fonctions à quatre paramètres donnée par la relation (2.9) et effectuer une optimisation, sous les contraintes $x \geq 0$, $y \geq 0$, $X \geq 0$, $Y \geq 0$, en minimisant l'écart entre les trois sorties z_1, z_2 et z_3 du système "pathologique" contrôlé (3.1) et les trois sorties φ_1, φ_2 et φ_3 du système "physiologique" (3.1) soumis aux entrées $e_1 = 0$, $e_2 = 0$ et $e_3 = p(t)$.

Ceci fera l'objet d'une prochaine étude, mais les simulations effectuées avec les entrées e_1 et e_2 sous la forme (3.3) (cf. [8]) montrent déjà qu'une meilleure approche de la courbe glycémique est obtenue avec l'intervention simultanée des deux actions X et Y (insuline et glucagon) plutôt qu'avec l'insuline seule.

IV. CONCLUSION

Nous avons présenté et illustré par deux exemples une nouvelle méthode de recherche liant étroitement l'automatique et la biologie. Cette voie dont Bernard-Weil est l'initiateur, ouvre un champ d'investigation immense en permettant, par un procédé de modélisation original qui s'apparente aux "dynamical metaphors" de Rosen [21] , de prenore en compte l'aspect ago-antagoniste qui intervient dans un grand nombre de régulations biologiques. Cette modélisation, à même de simuler aussi bien le pathologique que le physiologique, propose des contrôles bipolaires aux incidences thérapeutiques parfois surprenantes. Il n'est pas question que l'automaticien rentre dans les précisions médicales dont il n'a pas la compétence, mais il peut quand même indiquer, comme le montre déjà un certain nombre de publications médicales (cf. [5, 10, 11]), que la pratique des thérapeutiques bipolaires étend pas à pas son champ d'application. Il n'est pas douteux que dans un avenir que l'on doit rendre aussi proche que possible, ces thérapeutiques conduisent à supprimer l'état de souffrance d'un grand nombre d'êtres humains.

BIBLIOGRAPHIE

[1] M.S. BAZARAA et C.M. SHETTY, Nonlinear Programming, Theory and Algorithms, Wiley, New York, 1979.

[2] M. BOSHERNITZAN, Universal formulae and universal differential equations, Annals of Mathematics, 124, 1986, pp. 273-291.

[3] E. BERNARD-WEIL, Effects of a week of ACTH or corticosteroid treatment on the neuropostpituitary response to corticosteroid load, Steroids Lip. Res., 3 , 1972, pp.24-29.

[4] E. BERNARD-WEIL, Formalisation et contrôle du système endocrinien surréno-posthypophysaire par le modèle mathématique de la régulation des couples ago-antagonistes, Thèse d'Etat, Universté Paris VI, France, 1979.

[5] E. BERNARD-WEIL, Lack of response to a drug : a system theory approach, Kybernetes, 14, 1985,pp. 25-30.

[6] E. BERNARD-WEIL, Interactions entre les modèles empirique et mathématique dans la vasopressino-corticothérapie de certaines affections cancéreuses dans "Régulations physiologiques : Modèles récents", G. Chauvet et J.A. Jacquez, éd., Masson, Paris, 1986, pp. 133-155.

[7] E. BERNARD-WEIL, A general model for the simulation of balance, imbalance and control by agonistic-antagonistic biological couples, Mathem. Modelling, 7, 1986, pp. 1587-1600.

[8] E. BERNARD-WEIL et D. CLAUDE, Simulation du test de tolérance au glucose par le modèle de la régulation des couples ago-antagonistes. Contrôle bipolaire, C.R. Acad. Sci. Paris, 305, série I, 1987, pp. 303-306.

[9] E. BERNARD-WEIL, C. DALAGE, C. PIETTE et L. OLIVIER, Action of lysine-vasopressin on the protein content of Hela cell cultures and on the RNA and DNA concentrations of tissue incubation, Experta Med. Internat. Congr. Ser. : Protein and Polypeptide Hormones, 161, 1968, pp. 547-548.

[10] E. BERNARD-WEIL et B. PERTUISET, Mathematical model for hormonal therapy (vasopressin, corticoids) in cerebral collapse and malignant tumors of the brain (36 cases), Neurol. Res., 5, 1983, pp. 19-35.

[11] E. BERNARD-WEIL, J.L. JOST et P. VAYRE, Nouvel aspect des relations hôte-tumeur. Etude du système surréno-post-hypophysaire chez le cancéreux digestif, Chirurgie, 113, 1987, pp. 293-298.

[12] D. CLAUDE, Découplage des systèmes non linéaires, séries génératrices non commutatives et algèbres de Lie, SIAM J. Control Optimiz., 24, 1986, pp. 562-578.

[13] D. CLAUDE, Everything you always wanted to know about linearization but were afraid to ask, in "Algebraic and geometric methods in nonlinear control theory", M. Fliess and M. Hazewinkel, ed., D. Reidel Publishing Company, 1986, pp. 181-226.

[14] D. CLAUDE, Découplage et linéarisation des systèmes non linéaires par bouclages statiques, Thèse d'Etat, Université Paris-Sud, France, 1986.

[15] D. CLAUDE et E. BERNARD-WEIL, Découplage et immersion d'un modèle neuro-endocrinien, C.R. Acad. Sci. Paris, 299, série I, 1984, pp.129-132.

[16] M. FLIESS, Automatique et corps différentiels, Forum mathématiques, 1, 1989.

[17] U. KOTTA, Application of inverse system for linearization and decoupling, Systems Control Lett., 8, 1987, pp. 453-457.

[18] M. MONACO, P.H. KOHN, W.R. KIDWELL, J.S. STROBL et M.E. LIPPMAN, Vasopressin action on WRK-1 rat mammory tumor cells, J.N.C.I., 68, 1982, pp. 267-270.

[19] M. PAWLIKOWSKI, The effect of neuropeptides on cellular proliferation, Materia Medica Polona,Fasc. 1, 61, 1987, pp. 17-20.

[20] E. ROZENGURT, A. LEGG et P. PETTICAN, Vasopressin stimulation of mouse 3T3 cell growth, Proc. Natl. Acad. Sci. USA, 76, 1979, pp. 1284-1287.

[21] R. ROSEN, Dynamical System Theory in Biology, Wiley-Interscience, New York, 1970.

[22] A. RUBEL, A universal differential equation, Bull. Amer. Math. Soc. (N.S.), 4, 1981, pp. 345-349.

[23] A. SILVETTE et S. BRITTON, A theory of corticoadrenal and postpituitary influences on the kidney, Science, 88, 1938, pp. 150-151.

[24] M.K. SUNDARESHAN et R.A. FUNDAKOWSKI, Stability and control of a class of compartmental systems with application to cell proliferation and cancer therapy, IEEE Trans. Automat. Contr., AC-31, 1986, pp. 1022-1032.

[25] G.W. SWAN, Application of Optimal Control Theory in Biomedicine, Dekker, New York, 1984.

[26] A.T. WINFREE, The Geometry of Biological Time, Springer-Verlag, New York, 1980.

COMPUTER MODELS APPLIED TO CANCER RESEARCH

Werner Düchting
Department of Electrical Engineering
University of Siegen
Hölderlinstr. 3, D-5900 Siegen, West Germany

ABSTRACT: The aim of this contribution is to illustrate the impact of computer simulation in the field of biology and medicine. This paper shows how systems analysis, control theory and computer science can stimulate new approaches to interpret cancer, to predict tumor growth and to optimize tumor treatment.

Starting with a review of the current biological knowledge about the origin of cancer a computer model is constructed
- to simulate the time behaviour of disturbed cell growth control circuits
- to predict spatial tumor growth (2-D, 3-D) and
- to simulate different kinds of cancer treatment (surgery, radiation- and chemotherapy).

In the long run the aim of our work is to optimize treatment strategies and schedules in vitro and in vivo by computer simulation prior to clinical therapy.

1. BIOLOGICAL BACKGROUND OF THE CANCER PROBLEM

Cancer is a multistep process with the stages of initiation, promotion and progression. Characteristic features of malignant tumors (1) are uncontrolled proliferation, invasion in adjacent normal tissue, metastases induced to other tissues via lymphatic channels and the ability to evade immune surveillance. Though cancer treatment is concentrated on a prevention of metastases (2) the central question in the background of research is: Which is the initiating event that is responsible for a stepwise transformation of a normal cell into a tumor cell? Recent investigations in the field of molecular biology have focussed on dominant cellular genes called "proto-oncogenes" which can be activated by tumor viruses, gene amplification, gene translocation and genetic mutation. In spite of this progress (3) the main question how genes and the growth of normal and malignant cells are regulated still remains open.

Most of the normal tissues in the body contain some cells that can renew themselves (neurons, liver cells, kidney cells) if a tissue is injured. The division of a cell into two new ones involves four stages: G1 → S → G2 → M (G1 is a gap after stimulation; S is the phase of DNA replication; G2 is a second gap period and M is the stage of mitosis). When the replacement has been completed the repair process stops. Furthermore, at particular stages of the cell cycle the cells may be blocked by drugs or agents, or they may move out of the cell cycle into a resting phase known as G0 (4).

In contrast to the normal cell a tumor cell is theoretically able to divide indefinitely. In addition a different morphology, larger nucleus, abnormal number of chromosomes and the formation of new capillaries (tumor angiogenesis) which is associated with a more rapidly growing tumor (5) can be noticed.

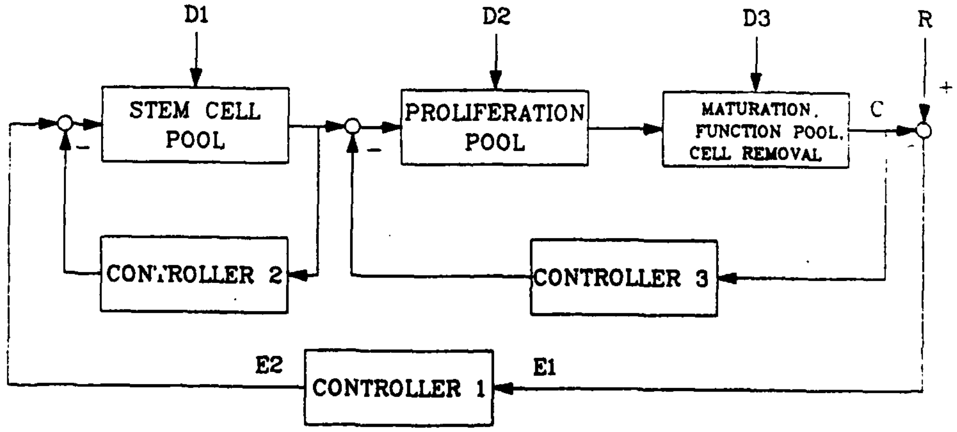
For studying the process of carcinogenesis tumors are induced to animals or to cell cultures (in vitro). Cell cultures are not only used to study the division of tumor cells, but also to determine the effect of chemotherapeutic drugs. During the last years a large progress has been made in experiments gaining hard data about normal and abnormal cell-growth control processes for instance of cell-cycle phase durations.

2. MODELING APPROACHES

Starting from basic biological test results a large body of mathematically oriented work applying mathematics to the field of biology and medicine has been published (6-10). Unfortunately these models which consist of complicated formulae, are in most cases not completely understood by clinicians. In this dilemma the combined application of methods of systems analysis, control theory, automata theory, computer sciences and heuristics is a good link between the diverging areas of medicine and mathematics.

Our own approach developing closed-loop control circuits for tumor growth started in 1968 (11). At that time the subject of consideration was focussed on stability conditions and on the interpretation of cancer as an unstable closed-loop control circuit. Step by step the dynamic behaviour of cell renewal control loops (Fig. 1) was investigated. Blockoriented simulation languages have been used for simulating the macromodels. As a result the number of cells as a function of time has been plotted (12).

Then oncologists advised us to consider not only the time but also the spatial behaviour of tumor growth. In a first approach we developed models at a cellular level which described the 2-D behaviour of a normal cell inoculated into a nutrient medium (in a Petri dish). Next we extended this approach and tried to simulate tumor growth in the tissue of a tobacco leaf (13).



R: Required tissue oxygen (desired number of erythrocytes)
C: Number of red blood cells (erythrocytes)
E2: Production of the erythropoietin hormone
D1, D2, D3: Disturbance

Fig. 1: Multi-loop control circuit of erythropoiesis

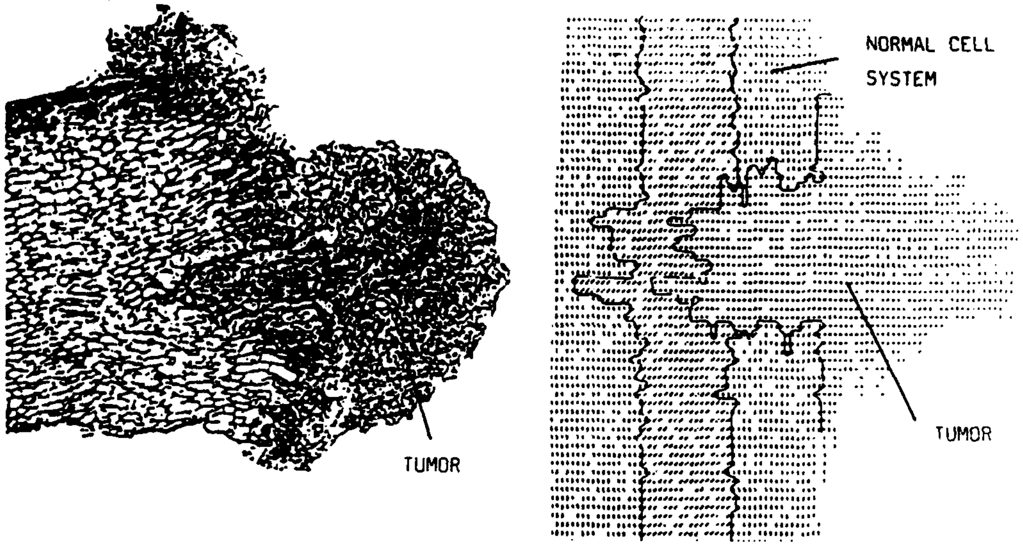


Fig. 2: Simulation of tumor formation in the tissue of a tobacco leaf

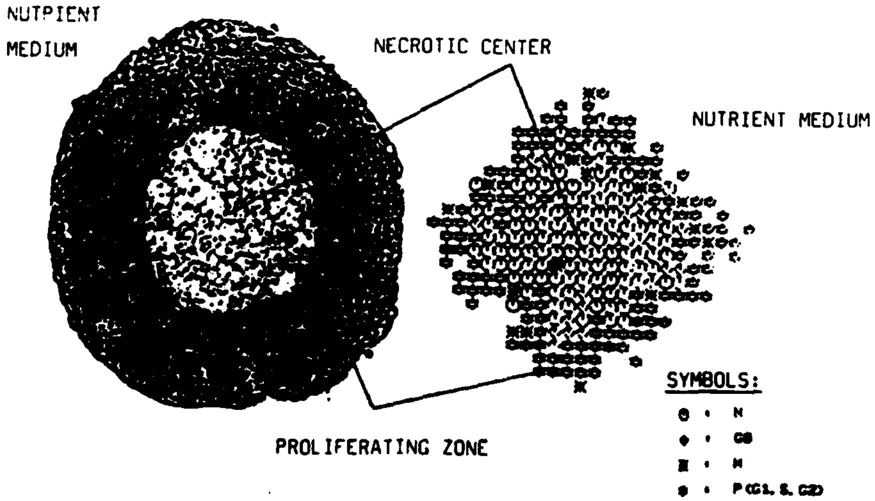


Fig. 3: Simulation of the formation of a tumor spheroid. The initial configuration consisted of a single tumor cell placed in the center of the nutrient medium

After getting the results shown in Figure 2 we improved these models by introducing distinguished cell cycle phases (G1, S, G2, M, GO, N). Thus, we were able to simulate the 3-D growth of a single dividing tumor cell (14) inoculated into the center of the cell space of a nutrient medium at the beginning of the simulation run (Fig. 3).

The introduction of distinguished cell-cycle phases was necessary because chemotherapeutic agents and rays effect only a very particular phase of the cell cycle that means they act phase - specifically.

After simulating in vitro tumor growth the attempt was made to substitute the nutrient medium by static blood vessels (15). However, very soon it was clear that a more realistic structure of capillaries was desirable for simulating in vivo tumor growth.

3. DESIGN STRATEGIES OF A HEURISTIC MODEL

The modeling of complex cell growth requires a considerable simplification. Some of the oversimplifying assumptions are

- constant volume of a cubic cell
- constant phase duration and constant cell loss
- only horizontal and vertical communication between neighboring cells
- a limited tissue volume by computer facilities
- side effects, immunologic reactions, heterogenity, drug resistance and the formation of metastases are neglected.

If you want to construct a model of high order, it is necessary to design a modular concept. In this case it means to design modular structured subsystems.

(i) You need cont~ol models (Fig. 4) which describe the cell division of normal and tumor cells at a cellular level including experimentally gained data e.g. of cell-cycle phase durations.

(ii) Heuristic cell-production and interaction rules are required describing the cell-to-cell communication. For instance one rule of the catalogue may say:
All tumor cells residing at a distance larger than 100 µm from the capillaries after the next division step will enter the resting phase GO.

(iii) Cell movement is described by transport equations (diffusion-, Poisson-equation), that means we have to introduce into the model gradients for pressure and metabolic compounds.

(iv) To represent 2-D and 3-D simulation results computer-graphics software packages are necessary.

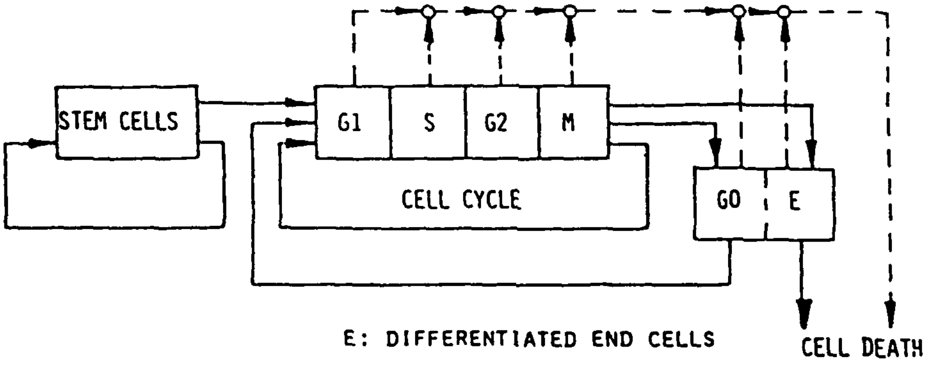


Fig. 4: Simplified cytokinetic model describing the division of a normal cell

The large body of statements, rules and equations has been transformed into algorithms. In addition algorithms considering tumor treatment (surgery, radiation- and chemotherapy) have been developed in subprograms written in FORTRAN IV. To start the simulation program packages the following input data have to be fed into the computer (VAX 730):

- notations about the character of a cell (normal, malignant)
- cell-cycle phase durations
- cell-loss rates
- initial configuration of normal tissue and of tumor cells
- distinguished data about the kind of the planned tumor treatment.

4. SELECTED SIMULATION RESULTS

Numerous simulation runs have been performed by Düchting and Vogelsaenger (15-17) simulating tumor growth and different kinds of treatment. Some special results will be demonstrated now.

4.1 Growth of capillaries

The simulation of in-vivo tumor growth requires a realistic structure of capillaries. Therefore Vogelsaenger (16) investigated the question: Is the formation of capillaries a stochastic or a regulated process? In (16) the assumption is made that each cell of an organ in evolution has a special request for oxygen and glucose. Therefore, parallel to the formation of tissue capillaries are built with a specific structure corresponding to the required oxygen and glucose. That means from the viewpoint of control theory the request for oxygen supply is regulated to a constant level by building a special structure of capillaries. A comparison between Figure 5 and Figure 6 shows that for the cortex of a rat the simulation result is highly similar to the experimental result received by Bär (18).

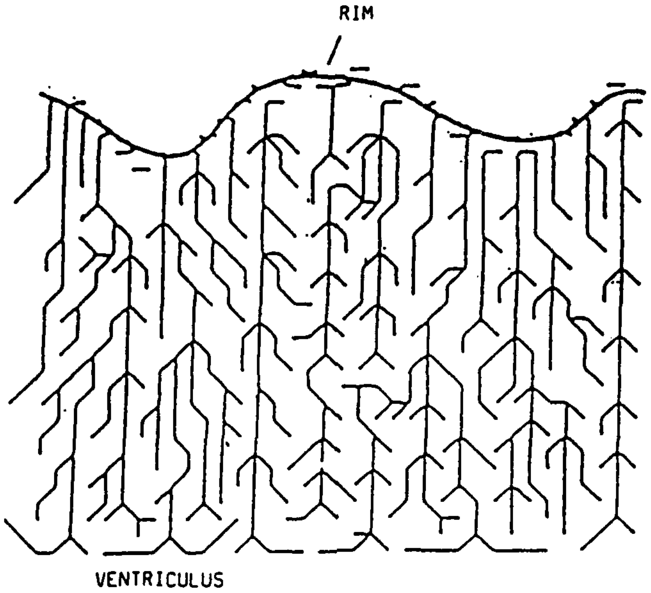


Fig. 5: Capillary network in the cortex (simulation result)

4.2 Spread of tumor cells in the cortex

Now the assumption is made that a single tumor cell is arbitrarily placed in the tissue of the cortex at T=1 unit of time. If this tumor cell resides close to a capillary it will divide and move in accordance with the cell production rules (Fig. 7). Further tumor growth is possible only because tumor cells produce a substance which is called tumor-angiogenesis factor (TAF). This factor stimulates nearby blood vessels to send out new capillaries (Fig. 8) which grow towards the tumor, penetrate it and lead to further rapid tumor growth. Recently great efforts have been made to attack cancer by trying to find a protein which inhibits the production of the tumor-angiogenesis factor.

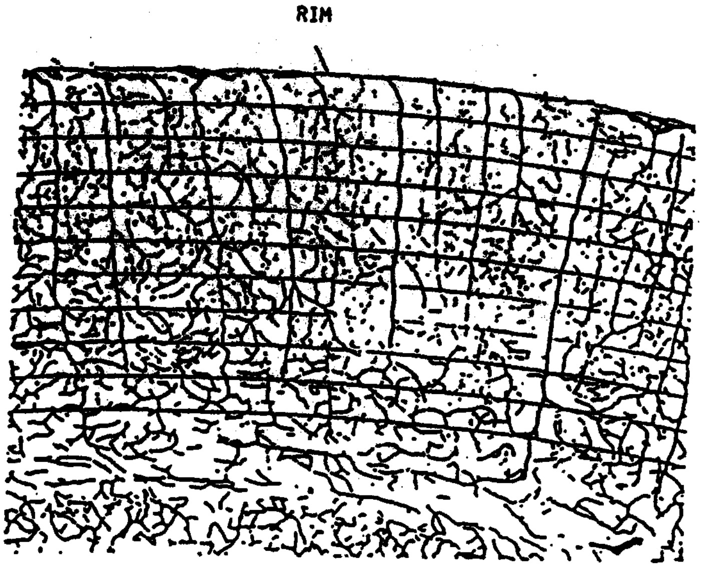


Fig. 6: Vascularization of the cortex (18)

4.3 Chemotherapeutic treatment in vitro

As pointed out in section 1, the cytotoxic effect of chemotherapeutic drugs is tested in cell cultures. These are very good invitro systems which can be simulated by a computer model. Figure 9(a) shows a tumor spheroid at T=200 units of time which has grown up from a single tumor cell inoculated into the center of the cell space at T=1 unit of time.

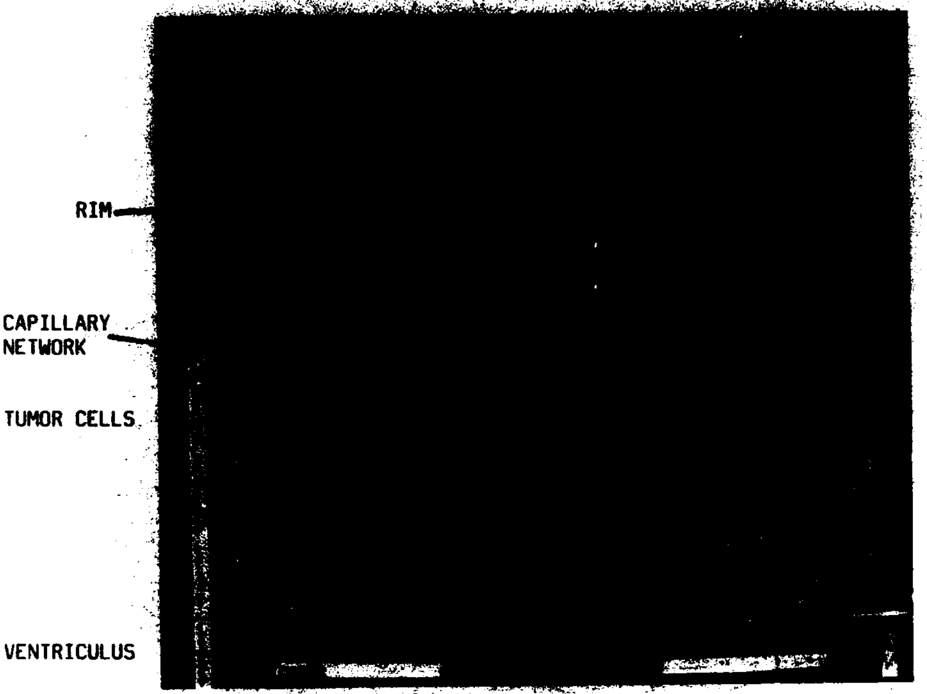


Fig. 7: Spread of tumor cells in the cortex at T=45 units of time

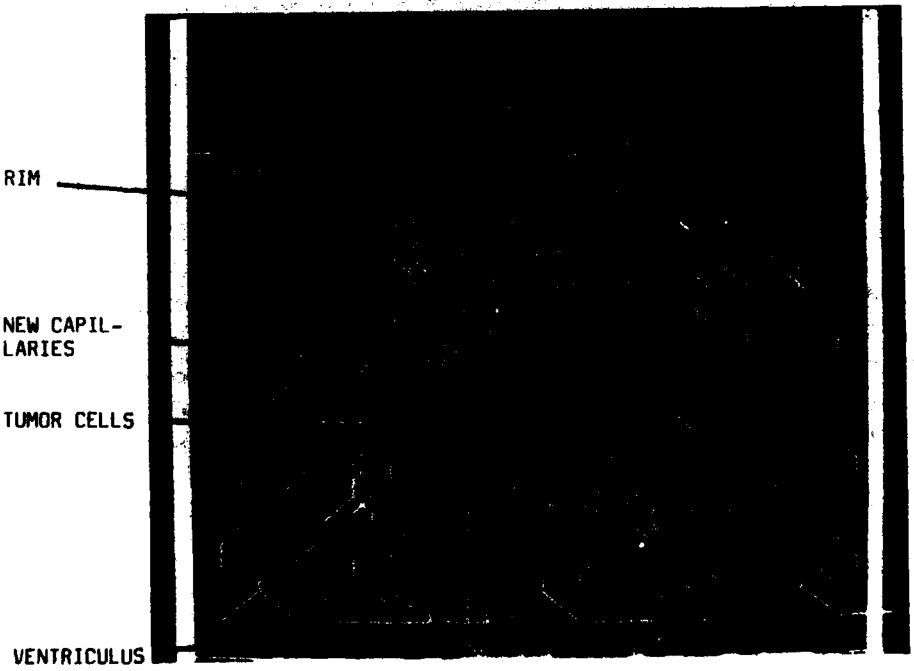
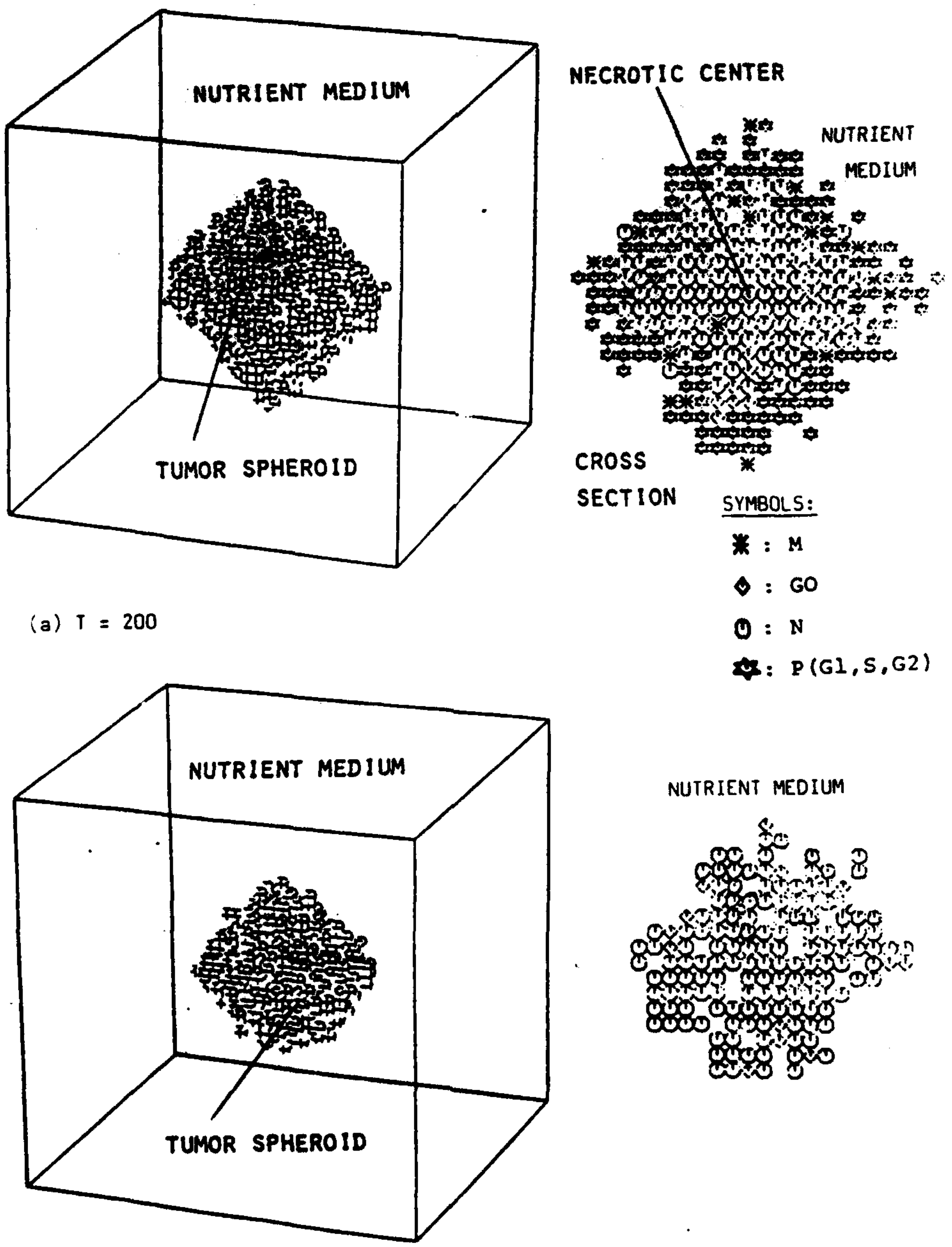
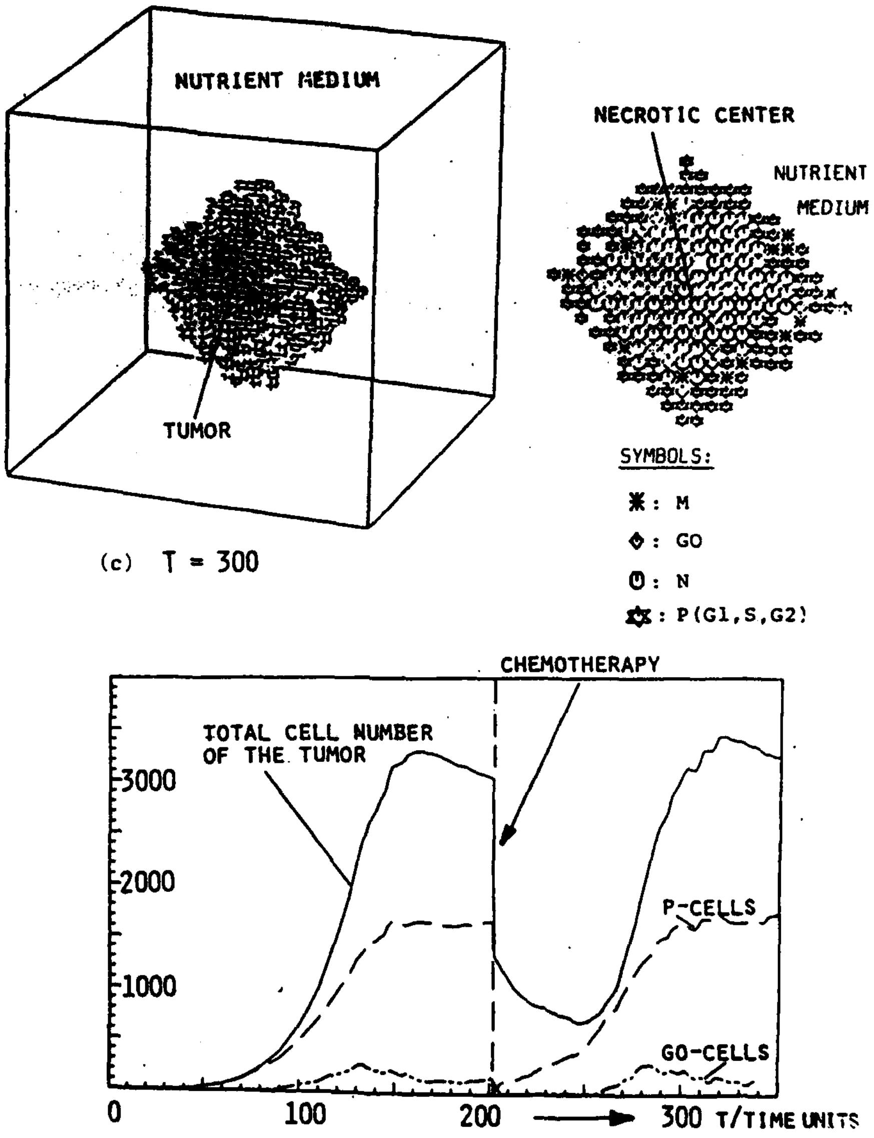


Fig. 8: Formation of new capillaries at T=120 units of time (tumor-angiogenesis effect)


NUTRIENT MEDIUM
TUMOR SPHEROID
NECROTIC CENTER
NUTRIENT MEDIUM
CROSS SECTION
SYMBOLS:
✱ : M
◆ : GO
⓪ : N
✱ : P(G1,S,G2)
NUTRIENT MEDIUM
TUMOR SPHEROID
NUTRIENT MEDIUM

(a) T = 200

(b) T = 201



(c) T = 300

(d) Number of tumor cells as a function of time

Fig. 9 (a)-(d): Simulation of a chemotherapeutic treatment of a tumor spheroid (in vitro)

At T=201 units of time it is assumed that all proliferating tumor cells (i.e. the outside rim) have been killed by a cytotoxic drug (Fig. 9(b)). Now the remaining resting tumor cells (G0-phase) in the neighborhood of the nutrient medium are being recruited into the cell cycle again, and after a short time of remission the tumor spheroid continues to grow (Fig. 9(c)-(d)). Therefore, a second therapeutic attack or a combined approach is recommended. The task which has been solved in (15) is to determine the optimum time at which the drug has to be applied for a second (and more) time(s).

5. FUTURE PROSPECTS

From the voluminous catalogue of unsolved problems in the area of cancer research I think there are three promising avenues of future work in the modeling field:

- Optimization of distinguished methods and schedules of cancer treatment.
- Generation of a more realistic initial configuration of a tumor by combining CT-pictures (Computer Tomography) with predictive models describing tumor growth and last not least
- Consideration of facts which had to be neglected so far (formation of metastases, immunologic reactions, drug resistance, heterogenity, side effects).

6. REFERENCES

(1) Tannock, I.F. and Hill, R.P. (eds.), The Basic Science of Oncology, Pergamon Press, New York 1987.

(2) Sherbet, G.V., The Metastatic Spread of Cancer, MacMillan Press, London 1987.

(3) Poste, G. and Crooke, St.T., New Frontiers in the Study of Gene Functions, Plenum Press, New York 1987.

(4) Baserga, R., Molecular Biology of the Cell Cycle, Int. J. Radiat. Biol., Vol. 49, No. 2 (1986): 219-226.

(5) Folkman, J. and Klagsbrun, M., Angiogenic Factors, Science, Vol. 235 (1987): 442-447.

(6) Cherruault, Y., Mathematical Modelling in Biomedicine, D. Reidel Publishing Company, Dordrecht 1986.

(7) Segel, L., Modeling Dynamic Phenomena in Molecular and Cellular Biology, Cambridge University Press, Cambridge 1984.

(8) Swan, G.W., Applications of Optimal Control Theory in Biomedicine, Marcel Dekker Inc., New York 1984.

(9) Wolfram, St., Cellular Automata as Models of Complexity, Nature, Vol. 311, No. 5985 (1984): 419-424.

(10) Meinhard, H., Models of Biological Pattern Formation, Academic Press, London 1982.

(11) Düchting, W., Krebs, ein instabiler Regelkreis, Versuch einer Systemanalyse, Kybernetik, 5. Band, 2. Heft (1968): 70-77.

(12) Düchting, W., Computer Simulation of Abnormal Erythropiesis - an Example of Cell Renewal Regulating Systems, Biomed. Techn. 21 (1976): 34-43.

(13) Düchting, W. and Dehl, G., Spatial Structure of Tumor Growth: A Simulation Study, IEEE Transactions on Systems, Man and Cybernetics SMC-10, No. 6(1980): 292-296.

(14) Düchting, W. and Vogelsaenger, Th., Three-Dimensional Pattern Generation applied to Spheroidal Tumor Growth in a Nutrient Medium, Int. J. Bio-Medical Computing 12(1981): 377-392.

(15) Düchting, W. and Vogelsaenger, Th., Aspects of Modelling and Simulating Tumor Growth and Treatment, J. Cancer Res. Clin. Oncol. 105(1983): 1-12.

(16) Vogelsaenger, Th., Modellbildung und Simulation von Regelungsmechanismen wachsender Blutgefäßstrukturen in normalen Geweben und malignen Tumoren, Dissertation Siegen, 1986.

(17) Düchting, W., Simulation of 3-D Tumor Growth and Radiation Therapy in "Proceedings of the International Symposium Computer Assisted Radiology" edited by H.U. Lemke, M.L. Rhodes, C.C. Jaffee and R. Felix, Springer-Verlag, Berlin 1987: 335-339.

(18) Bär, Th., Patterns of Vascularization in the Developing Cerebral-Cortex, CIBA Found. Symp. 100(1983): 20-36.

BIOLOGICAL SYSTEM RESPONSE PREDICTION BY APPLICATION OF STRUCTURE--ACTIVITY MODEL

Borka Jerman-Blažič, Irena Fabič-Petrač
Institut Jožef Stefan
Jamova 39, 61 111 Ljubljana, Yugoslavia
and Milan Randić
Dept. of Mathematics and Computer Science, Drake University Des Moines and Ames Laboratory-DOE, Iowa State University, Ames Iowa 50011, USA

ABSTRACT

The structure-activity models are primarily oriented towards the evaluation of molecular similarity. The approach and the model of structure-activity relations presented in this report is based on the similarity parameters developed by use of probability functions. The molecular structures are encoded as sequences of numbers representing counts of paths of different lengths. The similarity index between two compounds is calculated as the difference between the gains of information derived through the comparison of the corresponding molecular path sequences. The similarity index is used as a basic information for modelling the property prediction model. The correspondences between the ranks representing orderings according to the similarity index value are then searched and expressed as correlation indices. The correlation matrix represents the source of data for clusterisation of the compounds. Optimal classification is obtained after several testings with different threshold values. The classification of a compound with unknown biological activity into one of the obtained clusters of compounds with known biological responses represents the source of data for prediction procedure. The method is illustrated on a group of benzamidines.

1. Introduction

The biological response prediction models are often based on cluster analysis methods. Cluster analysis is a multivariate technique that identifies groups or clusters of related objects in a multidimensional space [1,2]. The classification is aimed towards the search of pattern points which correspond to natural and useful groups of chemical compounds. The location of pattern point within a cluster is used for semi-quantitative determination of biological activity [3] or other physicochemical property. This is an usual procedure for property prediction of non-available or not yet synthetized compounds.

The last decade brought in the chemical literature many different classifying algorithms. As a rule the results after the application of different algorithms to the same data set differ. Consequently, the choice of the classification algorithm must be done very carefully according to the nature of the studied problem. In general, all methods for identifying clusters in a multidimensional space contain some heuristic and arbitrary elements. Quantitative evaluation of the accuracy of the classification method and the prediction power is possible only in the case where the chemical data are abundant.

Basic assumption used in development of structure-activity models is the expectation that molecules with similar structural features will exhibit similar physicochemical properties and biological or pharmacological activities [4]. Structural similarity or dissimilarity of drugs finds application in quantitative structure-activity relationship (QSAR) studies and in drug design [5]. The definition of the similarity within the models is based on mathematical terms, which describe the chemical structure of the drug. The most difficult problem in modelling is the derivation of mathematical expressions for chemical structure encoding. The mathematical terms used in the model are expected to contain a lot of information and to have general applicability to different chemical systems under a variety of conditions.

The model presented in this report makes use of the mathematical property of the molecular graphs. The classification approach is based on the comparison of all possible molecule rankings (represented as strings) generated according to the similarity of a particular molecule from the initial set. The results of string comparison are expressed as correlation coefficients and used for group generation. The properties associated with a group are used as source data for prediction of an unknown biological response.

2. The developed model

2.1. Definition of the molecular similarity measure

An important problem in modelling structure-activity relations (SAR) is the definiton of the molecular similarity. The similarity itself is a mathematical relation with transitive, reflectional and equivalence properties, but the molecular similarity derived from the mathematical properties of the molecules does not always reflects in the same manner these mathematical relations in the real chemical world. In the real word there are other elements besides the chemical structure that govern the molecular behaviour.

The selection of appropriate molecular descriptors in SAR is of a great importance. In the chemical literature it is commonly accepted that major factors that govern the chemical events and biological activity are the molecular shape and the molecular structure [6]. In our model the molecular descriptors are derived from the molecular graph, not from the molecular physico-chemical properties [7]. We decided to use as mathematical descriptors a set of structural parameters already found useful in the study of structure-activity relationships. The molecular model used is the structural formula in which the hydrogen atoms are supressed according to the widely accepted practice [8]. The hydrogen atoms are less essential for the chemical behaviour and their presence can be always deduced if required. The characterization of the chemical structure is done by enumeration of the self-avoiding walks or paths with different length in the hydrogen depressed graph. The use of path codes in the discussion of similarity versus property or activity has shown that shorter paths, in particular the paths of lengths two and three, reflect the physicochemical properties of the compound [9] while paths of longer length, which encode the presence of structural details at larger separations reflect the molecular shape and are of interest for study of the biological activity [10]. Derivation and computation of self-avoiding paths may be found elsewhere [11,12].

The approach and the model of structure-activity relations presented in this report is based on the similarity parametres developed by use of probability functions. The sequence of path numbers for a compound A_i may be viewed as a distribution of a particular property of the molecular graph [13]. The domination of the size effect which may obscure the analysis when molecules of different size are considered, may be avoided by normalization of the path sequences. The path sequences are normalized by dividing the entries in every sequence represented as a vector x_i (x_i = [x_{ij}], j=1,2,..,m; m is the number of the longest path) by the number of atoms in the structure. We denote the normalized vector as x_i^o.

Following Jeffrey [14] the similarity between two chemical structures A_i and A_k belonging to a studied data set can be defined as [15]:

$$I(A_i/A_k) = \sum_{j=1}^{m} (p_{ij}-p_{kj})\log_2(p_{ij}/p_{kj}) \qquad (1)$$

where p_{ij} is the probability that a randomly selected element of the sequence A_i will be found in the j-th group of elements, p_{kj} is defined analogously.

In the case of strings of path numbers characterizing a molecule, the probability that a randomly selected path is in the group of paths with length j will be:

$$p_{ij} = x^o_{ij}/(\sum_{j=1}^{m} x^o_{ij}) \qquad (2)$$

As p_{ij} is calculated from the elements which reflect the molecular features of A_i the quantity I(A_i/A_k) measures quantitatively the degree of similarity s_{ik} between the molecules A_i and A_k. The similarity matrix S for a particular set of compounds is obtained by calculating the similarity indices s_{ik} for every pair (A_i,A_k) in the data set.

The similarity index in the developed model is used as a basic information for property prediction. The studied compounds are ordered according to the similarity index value between the compound A_i and all others componds in the data set. The ranking is obtained with ordering of the structures by ascending values of the similarity indices [15]. After generation of all rankings in the data set a nxn matrix is obtained. The relationship between the compounds may be expressed quantitatively by use of different string comparison methods: trace, alignment and listing [16]. The calculated quantities give an information about the similarity between two compounds "derived" from the similarity relationships of both compounds to all elements present in the data set. In that way all individual structural characteristics present in the data set are fully considered.

The correspondence between two particular sequences is calculated by counting the number of identities in traces generated for these sequences [15]. A trace between two sequences consists of lines connecting the elements from both sequences. An element can have no more than one line and the lines must not cross each other. If the elements connected by a line are the same then the pair represents an identity, if they are different, the pair constitutes a substitution. The result of the comparison between the elements of the sequences A and B is expressed as the quantity W(A,B). W(A,B) represents the number of different identities found in all traces generated by comparison of two ranks and diminished by one [15]:

$$W(A,B) = \sum_{i=1}^{n} (c_i -1) \qquad (3)$$

where n is the number of the sequence elements; c_i is the number of identities in the trace i on the right of element i. Detailed description of the method of calculation of W(A,B) is given in [15]. The correlation coefficient r for two compounds A_i and A_j is computed according to the expression [13]:

$$r_{ij} = [2W(A_i,A_j)/n(n-1)] \qquad (4)$$

where n is the number of the string elements.

The correlation coefficient r_{ij} is 1 if the orderings of the sequence elements compared are identical and approaches zero if they are completely different. The correlation coefficients r_{ij} for all possible rankings in the data base form the correlation matrix R. An element of R gives a quantitative estimation of the similarity between the compounds A_i and A_j. Two similar compounds i.e., A_i, A_j will generate always a similar ordering of the rest of the data set elements and they will have very high r_{ij}. If the contrary is true, then the value of r_{ij} will be low.

2.2. The compound clustering

The developed clustering algorithm makes use of the correlation matrix R. The first step of the procedure is the search for the most correlated compounds within the matrix R. The searched compounds represent the kernels of the future clusters. Each kernel in the very beginning contains only two compounds with the highest found r_{ij}. Other kernel elements are added according to the first threshold value defired on the base of the first chosen value of r_{ij}. The kernel elements are represented in Fig.1 with the sign @ (note that this sign represents two compounds, the compound i and j). The second step completes the clustering procedure. Another threshold value is defined for classification of the rest of the data set components. An element k is added to a particular kernel if the value of r for this element i.e., r_{kn} is bigger or equal to the second threshold value. If two or more kernels satisfy this condition then the element is added to the kernel with the highest mean value of r_{kn}. The second threshold value is lower than the threshold value for kernel generation. The prescribed values r in both steps may be changed during the clustering procedure according to the nature of the classified data. Sometimes, the first threshold value happens to be too high. This results in a small number of clusters. In this case the threshold value is decreased. On the other side, the criteria for this threshold value has to be high enough because if it is low the similarity criteria can be lost. Optimal classification of the compounds is obtained after several testings with different threshold values of the correlation index.

2.3. Property prediction

The membership information for a compound with unknown biological activity in a cluster with known biological responses is used for property prediction. The developed procedure for biological response prediction assumes that the values of the biological responses of all compounds in the cluster contribute to the value of biological response of the unknown compound. The contribution of a particular compound in the cluster is taken to be proportional to the degree of similarity of that compound to the unknown one. The property predicition procedure is as follows:

let compound x be classified into a group having n compounds. The average of the correlation coefficients of the group r_a (r_a is calculated as the sum of r_{ij} of the cluster divided by n) and the average of the biological responses for the group BR_a (BR_a is calculated as the sum of BR_j, $j=1,....,n$ divided by n) are used for prediction of the unknown value of BR by the application of the following equation:

$$BR_{predicted}(x) = BR_a + \sum_{i=1}^{n} ((BR_i - BR_a)(1+r_{xi} - r_a))/n \quad (5)$$

For a large data base calculation of the similarity index, the correlation coefficient as well as the property prediction requires a computer aid. A computer program has been developed in programming language Pascal and implemented on Vax 11/750. The procedure and the model is represented in Fig.2.

3. Applicatons

A group of compounds which consists of 73 benzamidines derivatives with dopamine receptor affinity has been taken as a basic data set. The experimental values of logP for these compounds have been taken from the work of Hansch and coworkers [17]. The obtained clustering consists of 11 groups, 12 compounds are not classified because of low values of their correlation coefficients. The correlation matrix is displayed in Fig.1 where eleven groups of higheły correlated compounds may be recognized. The values of logP presented on the left side of the figure are grouped very well too. The predicted values for selected compounds are shown in Table I. The thresholds values have been obtained after several attempts. Finally, as the optimal the kernel threshold value has been taken to be 0.94 and the cluster threshold value 0.8.

4. Discussion and conclusion

The method presented in this report shows that string comparison tehniques may be applied in chemical classification of compounds with similar biological activities. The developed method and models may be considered as an evidence how certain mathematical tehniques may be applied for derivation of the relationship between biological system response and structure of chemical compound i.e., the potential drug. The molecular path counts are found as suitable non-empirical parameters for description of the molecular structure. The same approach may be applied to other applications with other molecular descriptors and sequence comparison having different contents. Obtained predictions of biological responses are optimistic and suggest further development of the method.

5. References

1. Dubes, R.; Jain, K.J., Pattern Recognition 1979, 11, 235-254
2. Harrison, P.J., Applied Stat. 1968, 17, 226-236
3. Chu, C.K., Anal.Chem. 1974, 46, 1181-1187
4. Franke, R., Theoretical Drug Design Methods, Elsevier 1984, Amsterdam, p.12
5. Randić, M., Int.J.Quant.Chem:Quant.Biol.Symp. 1984, 11, 137-153
6. Trinajstić, N., Chemical Graph Theory, Vol.I,II, CRS Press, Boca Raton, Florida, p.18
7. Randić, M.; Wilkins, C.L., Theor.Chim.Acta 1980, 56, 45-51
8. Balaban, T.A., J.Chem.Inf.Comput.Sci. 1985, 25, 334-343
9. Randić, M.; Kraus, G.A.; Dżonova-Jerman-Blażič, B., Studies in Physical and Theoretical Chemistry 1983, 28, 192-205
10. Grossman, S.C.; Dżonova-Jerman-Blażič, B.; Randić, M., Int.J.Quant.Chem., Quant.Biol.Symp. 1986, 12, 123-139
11. Wilkins, C.L.; Randić, M., J.Chem.Inf.Comp.Sci. 1979, 19, 31-37
12. Randić, M.; Brissey, G.M.; Spencer, R.B.; Wilkins, C.L., Comp.&Chem. 1979, 3, 5-13
    Randić, M.; Brissey, G.M.; Spencer, R.B.; Wilkins, C.L., Comp.&Chem. 1980, 4, 27-32
13. Barysz, M.; Trinajstić, N.; Knop, J.V., Int.J.Quant.Chem: Quant.Chem.Symp. 1983, 17, 441-451
14. Jeffreys, H., Theory of Probability, 3rd., Claredon, Oxford, 1961, p.72
15. Jerman-Blażič, B.; Fabič, I.; Randić, M., J.Comp.Chem., 1986, 2, 176-188
16. Kruskal, J., SIAM Rev. 1983, 25, 201-237
17. Hansch, C.; Yoshimoto, M., J.Med.Chem. 1974, 17, 1160-1167

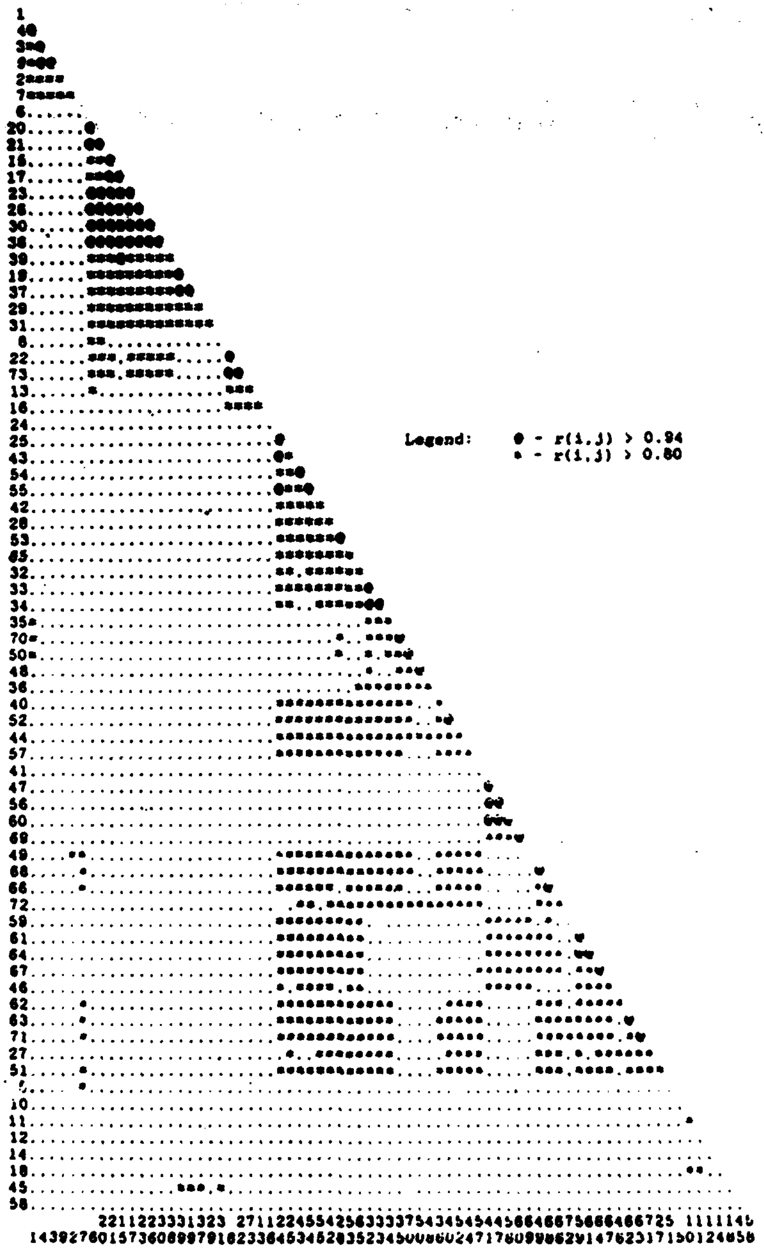


Fig.1. Clusterization of 73 benzamidine derivatives

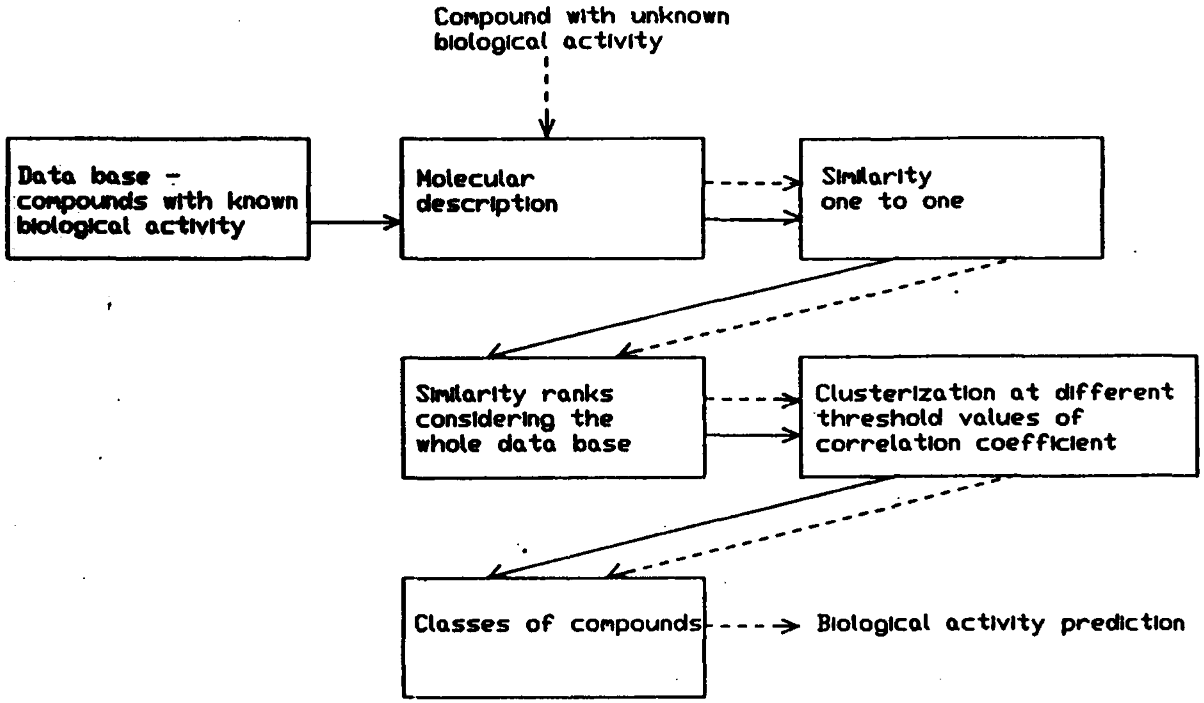

Compound with unknown biological activity
Data base - compounds with known biological activity
Molecular description
Similarity one to one
Similarity ranks considering the whole data base
Clusterization at different threshold values of correlation coefficient
Classes of compounds
Biological activity prediction

Fig.2. The biological activity prediction model

Table I. Predicted values of logP for benzamidine derivatives

Comp.No.	Measured BR	Calculated BR	Prediction error
3	2.35	2.47	5%
2	2.25	2.48	10%
8	2.68	3.10	16%
73	3.03	3.02	0%
24	3.77	4.11	9%
34	4.00	3.92	2%
40	4.09	4.22	3%
69	4.68	4.33	8%

CYCLIC CONTROL IN ECOSYSTEMS

Joseph Bentsman
Department of Mechanical and Industrial Engineering

Bruce Hannon
Department of Geography
Affiliate Scientist, Illinois Natural History Survey

University of Illinois at Urbana-Champaign
Urbana, IL 61801

The theory of feedback control as a possible stabilizing mechanism has already been introduced into ecosystem analysis. One problem in the theory is the identification of the informational links by which such controls operate. Cyclic controls, for example, zero–mean sine functions added to certain exchange flows in the system, might also contribute to system stability. Their advantage is that they operate without need for information from the rest of the system. The theory of ecosystem cyclic control is presented and applied to data from an oyster reef ecosystem.

I. INTRODUCTION

To address the problem of ecosystem stability and performance, the previous control studies utilized solely classical control principles, feedback and feedforward (Olsen, 1961; Lowes and Blackwell, 1975; Mulholland and Sims, 1976; Vincent, et.al., 1977; Goh, 1979; Hannon, 1985b,c, 1986; DeAngelis, 1986). If knowledge of the current output is used to modify the inputs to control the system, we have a feedback control situation (Wonham, 1984). Feedforward control uses current knowledge of the disturbance (rather than output) as the basis for a corrective action (Takahashi, et. al., 1970). The major problem with these kinds of controls, however, lies in explaining how the requisite information flows occur.

An alternative approach to ecosystem stability is found in the concept of cyclic (or vibrational) control (Meerkov, 1980; Bellman, et.al.,1986). Basically, cyclic controls are periodic variations (zero–mean) in the flows between components in an ecosystem or between the ecosystem and the surrounding environment. If the amplitudes and frequencies of these variations are within the appropriate range, the ecosystem, unstable without such variations, could under certain conditions be stabilized by their introduction without any information flows.

Oscillations–induced stabilization of ecosystems has been investigated by a number of researchers. Armstrong and McGehee (1976) developed a theory for the coexistence of a variety of species using a smaller number of resources. Their technique involved a the sequential staging of the species in a periodic manner, sharing the resource through time. Kemp and Mitsch (1979) used an empirical model to demonstrate the stable coexistence of three plankton species on the same resource if one of the resource inputs (wave energy) was regularly pulsing. They speculated that only a special range of frequencies and

Supported in part by the Illinois Department of Energy and Natural Resources

pulse amplitudes would produce the needed stability. The pulsing resource appeared to force a sharing between the three species, disadvantaging the species which was the most prolific under steady conditions. Levins (1979) established the sufficient conditions of coexistence by requiring that the resource or the species functions contain externally induced time-varying elements that enter the equations nonlinearly. Nonlinear dynamics in Levins' treatment was essential since it resulted in terms with even powers of zero mean oscillatory functions. The averages of such terms gave rise to the "average" nonzero inputs which acted as effective new resources and under certain conditions ensured stable oscillatory regimes of the system.

The goal of the present paper is to assess cyclic (vibrational) control theory as a tool in ecosystem analysis and management. We show that an unstable linear system can be made asymptotically stable by zero mean parametric excitations as well, and hence, nonlinearities are not necessary for oscillatory stabilization. We also utilize nonzero averages of even powers of zero mean oscillatory functions to obtain stabilizing corrections. However, we average not the original system with oscillations, but some other specially constructed system, the average of which reveals the dynamics of the original cycling system. For the purpose of illustration, we have chosen a modeling technique known as flow analysis (Hannon, 1973, 1985a; Barber, et. al., 1979) from a variety of ecosystem modeling approaches, each valid for certain system classes. First, we briefly review the flow analysis technique and present the theory of linear cyclic control of ecosystems. Then, we apply cyclic control to an oyster reef ecosystem where it acts in only one of the component flows. The extension of the theory to nonlinear systems can be done on the basis of the work of Bellman, et. al., 1986. The theory indicates the range of the amplitude/frequency ratios in which stabilizing cycles should be sought and asserts the existence of stabilizing cycles in this range. The actual stabilizing amplitudes and frequencies are determined via trial and error solutions of the differential equations.

II. FLOW ACCOUNTING

In the analysis of complex dynamic systems, it is necessary to develop consistent definitions and categorize all the identifiable flows. We start with the diagram shown in Figure 1. For more details on the ecological accounting system, see Hannon (1973), Finn (1976), Levine (1977, 1980), Hannon (1979), Patten, et. al., (1976), Herendeen (1981), Ulanowicz (1984) and Hannon (1985a).

In Figure 1, n x n matrix **P** is called the production-consumption matrix[1]. This matrix represents n processes which consume and produce n commodities. By process, we mean an aggregation of similar consumers-producers which is viewed as a single ecosystem component. By commodities, we mean the substances produced and consumed by the components of the ecosystem. The elements of the i^{th} column represent the breakdown of the main part of the consumption of the i^{th} process. The elements of the i^{th} row describe the breakdown of the main part of the production by the same process. Therefore, each element of **P** is the amount of commodity i (row number) which is used by process j (column number) in the given time period. For example, p_{ij} could be the daily amount of algal biomass (commodity i) consumed by a particular class of herbivores (process j). This is a multicommodity system since commodities listed along any of the rows are noncommensurable with commodities in any other row. Therefore, the row sums may be calculated since they are all the same commodity and, we assume, possess the same nutritional qualities for all consumers (The exception to this rule is the nonbasal heat of respiration which by definition has zero value to any component in the ecosystem). But, in general, the column sums cannot be formed because a common measure of a value of each element along the columns may not exist. Commodities of different qualities, even though measured in the same units (e.g.,gms-carbon) cannot be meaningfully added together. The inputs to omnivores and detritivores, for example, are of different qualities, both chemically and in nutritional meaning, to the consumer.

The diagonal elements in P are the self-use terms which are for example, own-waste consumption by rabbits and the consumption of decomposers by decomposers and cannibalism.

The full output vector q' is the sum of the vector of the nonbasal heat w given off by each of the components and the total output vector q.

The system in Figure 1 is shown without joint products, that is, each process (column) is assumed to produce a commodity of only one type. The joint product case is discussed in Hannon, 1985a and Costanza and Hannon, 1986.

The relationship to the external environment of the measurable quantities in the ecosystem modeled in Figure 1 is summarized in Table 1. The features of each quantity in this table are identified by the letters in the corresponding boxes. The table shows two vectors: r and e. The net output vector r is composed of three types of flows: exports (A & D), imports (D & E) and the heat of basal metabolism (B). By imports we mean those quantities which can be produced by the ecosystem but enter the system from the external environment. Exports are those quantities which can be produced by the ecosystem but which are not necessarily produced by it, and which leave the ecosystem for the external environment. The letter D in the import and export columns indicates those measured quantities which are passing through the ecosystem in the given time period, therefore, the quantity A – E is the net export. The system is perturbed by the externally induced change of the net export. The heat of basal metabolism (basal respiration) is that given off by the organism at rest. We take the heat of basal metabolism (B) as a surrogate for the commodity flows which are used in rebuilding the stocks metabolized during the given period. By stocks we mean the accumulated output quantities in each of the components in the system.

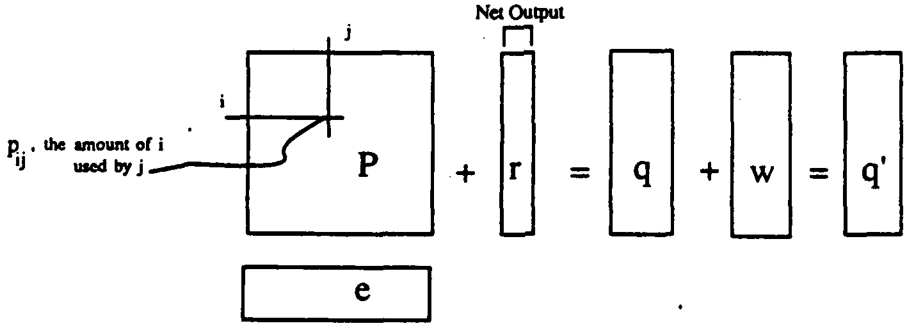


Figure 1. Steady State Ecosystem Flow Accounting Diagram

[1] Matrices are upper case symbols and vectors are lower case; both are in bold type. The elements of either are in plain type with the appropriate subscripting. A dot over a symbol indicates the time derivative.

			Net Output			Non-Produced Input
			r			e
			Exports	Imports	Basal Heat	
Commodities that the Ecosystem is capable of Producing	Produced by the System	Leaves the System	A		B	
		Stays				
	Not Produced by the System	Leaves	D	D		
		Stays		E		
Commodities that the Ecosystem is incapable of Producing						F

Table 1. Description of the Quantities which Form the Net Output and Nònproduced Input.

The stocks are, for example, the amount of biomass of algae which has accumulated in the producer (sun capturing) component of an aquatic ecosystem. The vector e stands for those input commodities that the ecosystem is incapable of producing (e.g., sunlight) but that are necessary for ecosystem functioning.

III. FLOW ANALYSIS

Next we combine the flow definitions above with the possibility of a growth in the stock of process j during the given time period Δt. These flows are graphically shown in Figure 2 for the individual process.

The consumption flows p_{ij}, production flows p_{jk} and the storage flow $\Delta s_j/\Delta t$ are internal to the ecosystem boundary, while the net output flows r_j, the nonbasal respiration flow w_j and the nonproduced input flow e_j cross the ecosystem boundary. The nonbasal respiration flow (e. g., the energy used in chasing prey, avoiding predators, food–searching and reproduction) is of such low quality that it cannot be utilized further by the ecosystem, and it is therefore considered a waste. The r_j consists of the net export of the process (export minus import) and the stock replacement (basal respiration). The net input vector e is assumed to cause no restriction to the level of q_j and is dropped from further consideration at the current stage of the model development.

The total outflow q'_j is defined for the steady state ecosystem as

$$q'_j \triangleq \sum_{k=1}^{n} p_{jk} + r_j + w_j \ . \tag{1}$$

To take into account a growth in stock, Δs_j, over the time period Δt when the system is not in the steady state, definition (1) is augmented as

$$q'_j \triangleq \sum_{k=1}^{n} p_{jk} + r_j + w_j + \frac{\Delta s_j}{\Delta t} \quad . \tag{2}$$

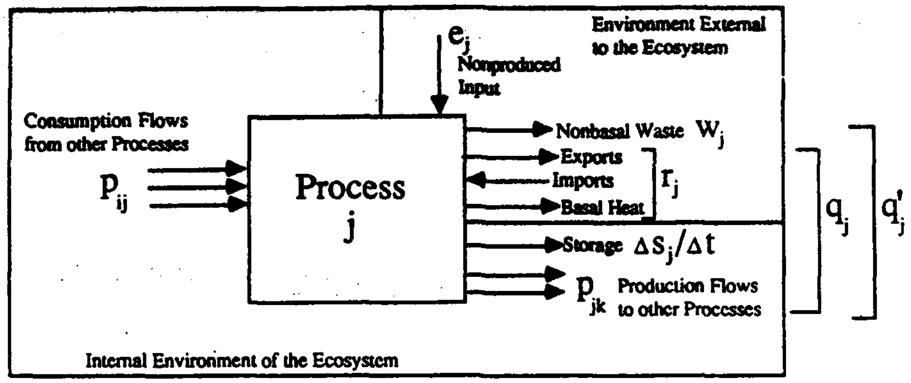

Environment External to the Ecosystem
e_j
Nonproduced Input
Consumption Flows from other Processes
p_{ij}
Process j
Nonbasal Waste W_j
Exports
Imports
Basal Heat
r_j
Storage $\Delta s_j/\Delta t$
p_{jk}
Production Flows to other Processes
q_j
q_j'
Internal Environment of the Ecosystem

Figure 2. The Definition of the Input and Output Flows of a Typical Process (j).

Three important simplifying assumptions are now made for the ecosystem shown in Figure 2 with q_j' defined in (2).

i) a commodity weighting or importance factor is assigned to each of the commodities produced in the system. The weight for each commodity is independent of which component consumes this commodity. A weight of zero is given to the nonbasal heat of respiration, and therefore, the vector w *disappears from the formulation. The element q_j can be then be formed* by the simple addition of all the elements along the j^th row of matrix **P**, the rate of the j^th stock growth and the j^th element of vector **r**. For a more complete discussion of the commodity weighting issue, see Hannon, 1985a.

ii) the inputs to process j, p_{ij} form a constant ratio with the output of process j, q_j. Thus, $p_{ij}/q_j = g_{ij}$ = constant. The constants g_{ij} are determined from the data on the ecosystem at its steady state and they are assumed to remain constant for the dynamic form of our model presented below. These constants represent the internal behavior of the j^th process. The g_{ij} incorporate the consumption flows into the model by locking them into a constant relationship with the output of the receiving process. Thus, the problem of summing the consumption flows (see Figure 2) is avoided.

iii) the stock (s_j) of any process (j) stays in constant proportion to the total output (q_j) of this process. That is: $b_{jj} = s_j/q_j$ = constant, forming a diagonal matrix $\mathbf{B} = \mathrm{diag}\{b_{11},\ldots,b_{nn}\}$. This assumption allows us to obtain a balance equation using definition (2) since now

$$q_j = s_j/b_{jj} \quad . \tag{3}$$

If the results of assumptions i) and ii) are combined with (2) and (3), and if Δt becomes infinitesimal, we have

$$\frac{s_j}{b_{jj}} = \sum_{k=1}^{n} g_{jk}\frac{s_k}{b_{kk}} + r_j + \dot{s}_j \quad , \tag{4}$$

where $\dot{s}_j = ds_j/dt$.

Equation (4) is the dynamic description of the stock for process j. However, most experimental ecosystem data is presented as flows. Therefore, we change (4) into a dynamic description of the flows for process j. Substituting (3) and its time derivative into (4) yields

$$b_{jj}\,\dot{q}_j = q_j - \sum_{k=1}^{n} g_{jk}\, q_k - r_j \quad \forall\, j,\ 1 \le j \le n,$$

or in matrix form

$$\dot{q} = Aq - B^{-1}r, \qquad A \triangleq B^{-1}(I - G). \qquad (5)$$

This time invariant ordinary differential equation (5) is in the "standard" form for the flow analysis approach.

IV. STABILITY ANALYSIS

The stability properties of the behavior of q when the system is subjected to a step change in r depend entirely on the matrix A in (5). If the real parts of all the eigenvalues of A were negative, the system would respond in a stable manner (Luenberger, 1979, p 158). However, in (5) the sum of the eigenvalues of matrix A is always positive. Therefore, the system will always respond to "sufficiently rich" changes in r in an unstable manner.

From an ecological viewpoint, a positive r represents an output of the ecosystem (for example, the amount of fish caught in the annual season). From the control theory viewpoint however, this output represents an *input* to the system or a *control* action. For example, the amount of fish caught directly affects the rate of (re)production of fish *and* many other quantities produced in the ecosystem, which in turn, also affect the fishing success. If the system (5) is to accurately represent the functioning of an ecosystem, the equations must be judiciously modified to include stabilizing or controlling flows. Equations (5) can be made to respond stably by modifying r to include a feedforward or a feedback control. Let us, however, demonstrate the use of cyclic control for ecosystem stabilization through the addition of a cyclic flow to one of the elements in the matrix G.

In the flow accounting framework, cyclic control alone cannot guarantee stability of the system. However, only a very simple form of constant feedback is required to make cyclic control effective. Such feedback can be easy to maintain since it need not ensure stability but only "condition" the system for cyclic control. On the other hand, for a broad class of the so-called decentralized systems, no constant or time-varying feedback exists that can stabilize the system (Anderson and Moore, 1981). In these cases, the addition of cyclic control can result in the desired stabilizing effect.

Since equation (5) is still always unstable, several changes must be made to r to demonstrate the cyclic control. First, r must be broken into two parts: a vector of net outputs which are independent of the output q, and another vector which contains the feedback and cyclic control and depends on q. The first vector contains the "set point" vector for the system, r_s: the vector of net outputs which in the absence of cyclic control determines the unstable steady state level q_s of the total output. The introduction of cyclic control converts the unstable steady state q_s into asymptotically stable T–periodic operating regime, $q_s(t)$, where T is the period of a cyclic control. A feedback control is needed to convert the trace of the matrix in equation (5) to a negative value (Meerkov, 1980). Assume that this is an internal control that changes the net output from the system in linear proportion to the production flows, a "flow" control (Hannon, 1986). For simplicity, let the linear proportionality be represented by a diagonal matrix of constants, Q. In this case, vector r in equation (5) is given by:

$$r = r_1 + r_s = r_s - r_c + Qq \quad ,$$

where $r_c = Qq_s(0)$

Equation (5) then becomes:

$$\begin{aligned}\dot{q} &= B^{-1}(I - G - Q)q + B^{-1}(r_c - r_s) \\ &= Nq + B^{-1}(r_c - r_s), \; N \triangleq B^{-1}(I - G - Q). \qquad (6)\end{aligned}$$

The constant vector $B^{-1}(r_c - r_s)$ will be dropped because it is independent of q and therefore does not affect the stability analysis.

Matrix Q must have only one non-zero element with sufficiently large absolute value to cause a sign change in the trace of N. Therefore, we further assume that matrix Q makes the trace of matrix N negative, but does not guarantee system stability, i. e., we simulate the circumstances where the feedback controls (like Q) are not adequate to make all of the eigenvalues fall in the left-half plane. This situation can arise if the information gathering processes of the system are somehow limited, resulting in lack of controllability and/or observability (Luenberger, 1979), but are sufficient to condition the system for cyclic control.

Let us again augment the vector r = Qq - D(t)q, where D(t) is a periodic, zero mean matrix. The periodic input D(t), is weighted by the state vector of the system q, and therefore D(t) appears in the system equation in the form of parametric perturbations or *cyclic control*. In this case equation (6) becomes

$$\dot{q} = [N + B^{-1}D(t)]q \; . \qquad (7)$$

Because equation (7) is time-varying, eigenvalues can no longer describe its stability. It is possible, however, to associate stability properties of the oscillatory system (7) with a certain constant matrix that describes its <u>average</u> behavior. The stabilizing action of cyclic controls consists in converting the remaining right-half plane eigenvalues of system (6) into "left-half plane on-the-average" ones. In this case, stabilization is achievable without the need for additional information flows, provided that the amplitudes and frequencies of the cyclic controls are within a critical range.

Assume, for simplicity, that the ij^th element of the cyclic control matrix D(t) is given by $d_{ij}(t) = c_{ij}\cos(\omega_{ij}t)$, where c_{ij} is the amplitude and ω_{ij} is the frequency of the oscillation.

In order to describe the average behavior of system (7), we introduce the parameter ε as

$$\varepsilon \triangleq \max_{ij}(1/\omega_{ij})$$

and define

$$c_{ij} \triangleq \alpha_{ij}/\varepsilon \quad \text{and} \quad \omega \triangleq \beta_{ij}/\varepsilon$$

so that the ij^th element of D(t) can be rewritten as $d_{ij}(t) = (\alpha_{ij}/\varepsilon)\cos(\beta_{ij}t/\varepsilon)$.

With this notation, the cyclic control matrix D(t) takes the form

$$\mathbf{D}(t) = \frac{1}{\varepsilon}\mathbf{D}'\left(\frac{t}{\varepsilon}\right),$$

and system (7) becomes

$$\dot{\mathbf{q}} = [\mathbf{N} + \frac{1}{\varepsilon}\mathbf{B}^{-1}\mathbf{D}'(\frac{t}{\varepsilon})]\mathbf{q}. \tag{8}$$

Thus, if the α_{ij}'s and β_{ij}'s are assumed constant, the amplitudes c_{ij} and the frequencies ω_{ij} of the zero-mean cyclic terms $d_{ij}(t)$ are parameterized by a positive ε. It has been proven (Bellman, Bentsman and Meerkov, 1985) that there exists an ε_0 = constant > 0, such that for any ε satisfying the inequality $0 < \varepsilon < \varepsilon_0$, the stability properties of system (8) are defined by the eigenvalues of a constant matrix

$$\mathbf{M} = \lim_{T\to\infty}\frac{1}{T}\int_0^T \Phi(\tau)^{-1}\mathbf{N}\,\Phi(\tau)\,d\tau \quad , \tag{9}$$

where $\Phi(t)$ is the state transition matrix of

$$\frac{d\mathbf{q}}{d\tau} = \mathbf{B}^{-1}\mathbf{D}'(\tau)\mathbf{q} \tag{10}$$

where $t = \tau/\varepsilon$.

Specifically, for sufficiently small ε, system (8) is asymptotically stable if all the eigenvalues of **M** have negative real parts. As seen from this result, the elements of matrix **M** are defined in terms of the elements of matrices N, $\mathbf{B}^{-1}$, "amplitude/frequency" ratios α_{ij}, and "frequency/frequency" ratios β_{ij}. Consequently, M provides a link between α_{ij}, β_{ij} and stability of (7): If α_{ij} and β_{ij} are found which place all the eigenvalues of M in the left-half plane, then there exists an ε such that oscillations with amplitudes α_{ij}/ε and frequencies β_{ij}/ε guarantee asymptotic stability of system (7). The matrix

$$\mathbf{M}' \triangleq \mathbf{M}\text{-}\mathbf{N}$$

can be thought of as a "correction" of N induced by oscillations.

In the context of ecological systems, cyclic control is easy to apply. Indeed, ecological systems are usually described by sparse matrices and therefore the cyclic control matrix D(t) might often satisfy condition $\mathbf{D}^2(t) = 0$ independently of the magnitudes and frequencies of the oscillations. In this case, since **B** is a diagonal matrix, all non-zero elements of matrix **M'** are given as

$$m'_{ij} = -\frac{\gamma_{ij}^2}{2}\,n_{ji} \quad , \quad \gamma_{ij} \triangleq \frac{1}{b_{ii}}\,\frac{\alpha_{ij}}{\beta_{ij}}, \tag{11}$$

where n_{ji} denotes the jith element of the matrix **N**. Therefore, the only elements of D(t) that will affect the eigenvalues of M are those off-diagonal elements that have a corresponding non-zero symmetric element in N.

The first step in the search for amplitudes and frequencies of the stabilizing oscillations is to find m'_{ij}'s that move all the eigenvalues of N+M' to the left-half plane. A straightforward way to accomplish this is to try only one of the appropriate elements at a time, and let g_{ij} increase from 0 to a sufficiently large number. When the appropriate set of elements m'_{ij}, and, hence, g_{ij} have been identified, we must

Table 2. Oyster reef Input-Output flow matrix (P), along with vectors for net export + stock replacement respiration (r), total output excluding waste heat (q), waste heat (w), and total output including waste heat (q').

	P									
	(1)	(2)	(3)	(4)	(5)	(6)	r	q	w	q'
Oysters 1	0	15.79	0	0	0	0.51	17.80	34.10	7.365	41.47
Detritus 2	0	0	8.17	7.27	0.64	0	6.19	22.27	0	22.27
Microbiota 3	0	0	0	1.21	1.21	0	2.875	5.295	2.875	8.17
Meiofauna 4	0	4.24	0	0	0.66	0	1.75	6.65	1.75	8.4
Deposit Feeders 5	0	1.91	0	0	0	0.17	0.215	2.295	0.215	2.51
Predators 6	0	0.33	0	0	0	0	0.2	0.53	0.15	0.68
Net Input e	41.47	0	0	0	0	0				
Control Q	1.52	2.28	.94	1.26	2.09	1.38				

return to equation (8), placing $(\alpha_{ij}/\varepsilon)\cos(\beta_{ij}/\varepsilon)$ at these locations in D(t). Then, by changing ε and repeatedly solving equation (8) for stabilizing pairs of (α,β), the areas of stabilizing amplitudes and frequencies can be found. The search for stabilizing oscillations becomes complicated when the stabilizing matrices that satisfy $D^2(t) = 0$ do not exist (see for example, Wu, 1975).

Cyclic control could naturally arise in an ecosystem as i) an oscillation of the flows between various components or ii) a part of the net output, a cyclical export (import) from (to) a particular component, the interpretation used in this paper.

What follows is a simple example of ecosystem stabilization by a cyclic control.

V. APPLICATION TO THE OYSTER REEF ECOSYSTEM

In this section, we apply the theory presented above to the oyster reef ecosystem (Dame and Patten, 1981). This compact but complex system is shown at steady state (i.e., for constant flows) in Figure 3.

The data from Figure 3 have been arranged in the proposed accounting framework (Figure 1) in Table 2. In this arrangement, estimates of the basal metabolism or structural-rebuilding respiration are included in the net output.

From the data in Table 2, we constructed G for use in the N matrix. With the feedback control elements of diagonal matrix Q, shown in Table 2, the trace of N is negative and its eigenvalues are: 0.0726 ±0.0371i, -0.1753, -.0089, -0.0994 and -0.0028. Because the complex pair has positive real parts, the system is unstable. Let us demonstrate that a cyclic control can be found to stabilize the system at the given steady state.

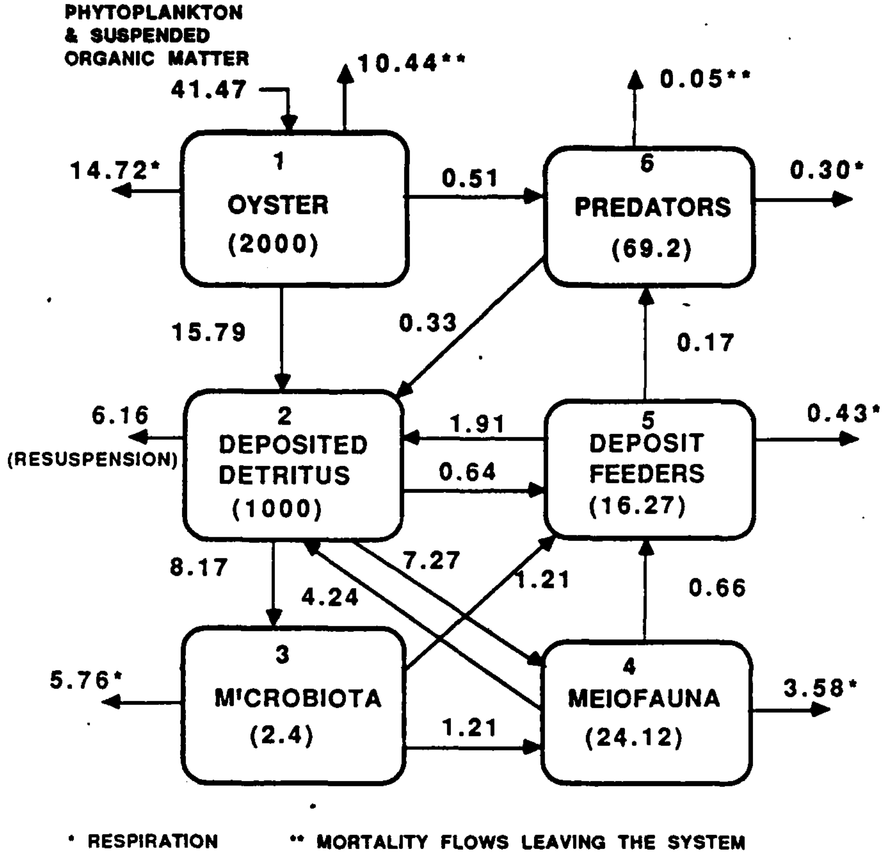


Figure 3. The Oyster Reef Ecosystem. Flow units are kcal/m –day.^{2} Stock unit: kcal/m^2.

Let $m'_{5,3}$ be the only non-zero element of matrix M', indicating a cyclic net input to deposit feeders and a cycle in the flow 5–3. Then by experiment, for $m'_{5,3} > 0.0346$, all the eigenvalues of matrix N + M' are in the left-half plane. Choosing $\beta_{5,3} = 1.0$, from equation 11 we obtain

$$\alpha_{5,3} = b_{5,5}\left[\frac{-2m'_{5,3}}{n_{3,5}}\right]^{\frac{1}{2}} = 1.7298,$$

where $b_{5,5} = 7.0893$ and $n_{3,5} = -1.1632$. Thus, according to the theory of Section IV, oscillations of the form $d_{5,3}(t) = \alpha_{5,3}\omega \mathrm{Sin}(\omega t)$, $\alpha_{5,3} > 1.73$, should stabilize the system for sufficiently large ω. The asymptotic nature of the theory implies however, that condition $\alpha_{5,3} > 1.73$ should be partially observed for smaller ω as well. It is precisely this insight that motivates the numerical search for the actual parameters of stabilizing cycles at low frequencies. In Figure 4, we demonstrate that condition, $\alpha_{5,3} > 1.73$, is partially observed for $\omega/2\pi > 0.08$. The amplitudes are $q_3 d_{5,3}(t)/b_{5,5}$. The cross-hatched region in Figure 4 corresponds to the actual stabilizing amplitudes and frequencies of the cycles $d_{5,3}(t)$.

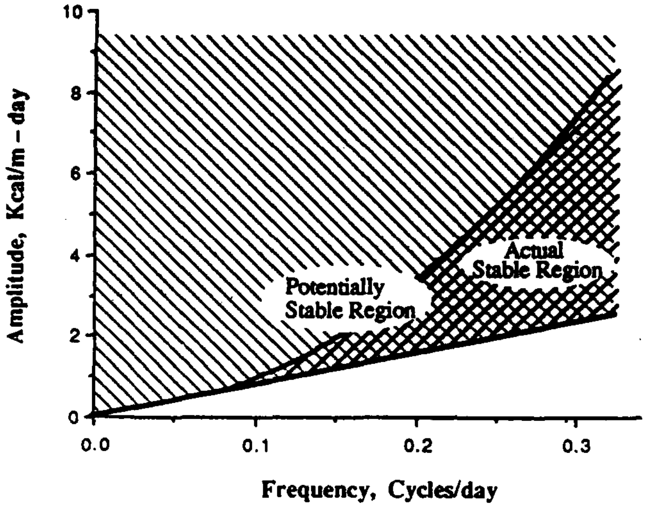


Figure 4. Cyclic Control in the Oyster Reef Ecosystem. The Range of the Parameters of the Stabilizing Oscillations of the Net Input to the Deposit Feeders 5 and of the Connection to the Microbiota 3.

While our choice of Q was largely arbitrary, we find the data in Figure 4 interesting. They show, for example, that a cyclic net input to the Deposit Feeders (which in turn allows them to cycle their feeding on the microbiota) can stabilize this ecosystem (given the above Q). With a cycle frequency of once in seven days, the stabilizing amplitude would range from about 1.1 to 1.7 kcals/m^2–day, encompassing the average value of the flow from 3 to 5 of 1.2 kcals/m^2–day (see Figure 3). It seems possible that such a cyclic flow could occur. No data on the variation of flows in this oyster reef ecosystem were given (Dame, 1976, 1979; Dame and Patten, 1981). From Figure 4, we also see that smaller stabilizing amplitudes are associated with lower frequencies. This application to the oyster reef system is expected to convey a biological possibility of ecosystem stabilization by already existing or intentionally introduced oscillations.

VI. CONCLUSION

The material presented above demonstrates that cyclic control is a biologically feasible stabilizing mechanism that could either develop in the course of evolution or be introduced by an ecosystem manager.

The important point about cyclic control is that stabilization can be provided without any information exchange. Therefore, the components that can establish a balanced cyclic exchange of materials or energy with the external environment and/or with other components might bring stability to the whole system without the cost of building and maintaining additonal information links. Thus, since cycles often occur in ecosystems naturally or can be introduced intentionally, cyclic control theory constitutes a viable tool for the ecosystem analysis and management.

ACKNOWLEDGMENT

The first author is grateful to his teachers and co-workers S.M. Meerkov and R.E. Bellman for arousing his interest in the application of control concepts to living systems. Both authors thank Salvatore Cusumano for his help in setting up the differential equation solver.

REFERENCES

Anderson, B. and J. Moore, 1981, Time Varying Feedback Laws for Decentralized Control, IEEE Trans. Autom. Control, AC–26, 5, 1133–1138.

Armstrong, R. and R. McGehee, 1976, Coexistence of Species Competing for Shared Resources, Theo. Population Biol., 9, 317–328.

Barber, M., B. Patten and Finn, J., 1979, Review and Evaluation of I-O Flow Analysis for Ecological Systems, in: Compartmental Analysis of Ecosystem Models, Vol. 10 of Statistical Ecology, Matis, J., Patten, B. and White, G., eds., International Cooperative Publishing House, Fairland, Md.

Bellman, R., J. Bentsman and S. M. Meerkov, 1985, Stability of Fast Periodic Systems, IEEE Trans. on Automatic Control, AC-30, 3, 289-291.

________, 1986, Vibrational Control of Nonlinear Systems, Vibrational Stabilizability, IEEE Trans. Autom. Control, AC–31, 8, 710–716.

________, 1986, Vibrational Control of Nonlinear Systems, Vibrational Controllability and Transient Behavior, IEEE Trans. Autom. Control, AC–31, 8, 717–724.

Costanza, R., C. Neill, S. Leibowitz, J. Fruci, L. Bahr, and J. Day, 1983, Ecological Models of the Mississippi Deltaic Plain Region: Data Collection and Presentation, U. S. Fish and Wildlife Service, Washington, DC., FWS/OBS–82/86.

Costanza, R. and B. Hannon, , 1987, Multicommodity Ecosystem Analysis: Dealing with Apples and Oranges in Flow and Compartmental Analysis, to appear in: Progress in Systems Ecology: Mid–1980's Issues and Perspectives, B. Patten and S. Jørgensen, Editors.

Dame, R. and B. Patten, 1981, Analysis of Energy Flows in an Intertidal Oyster Reef, Marine Ecology Progress Series, 5:115-24. See also: Dame, R., 1976, Energy Flow in An Intertidal Oyster Population, Estuarine and Coastal Marine Science, 4, 243-253. See also: (1979), The Abundance, Diversity and Biomass of Macrobenthos on North Inlet , South Carolina, Intertidal Oyster Reefs, Proceedings of the National Shellfisheries Association, 69, 6-10.

DeAngelis, D., W. Post and C. Travis, 1986, Positive Feedback in Natural Systems, Biomathematics, Springer–Verlag, Berlin, **15**, 233.

Finn, J. 1976, Measures of Ecosystem Structure and Function Derived from Analysis of Flows, J. Theo. Biol., **56**, 363-380.

Goh, B. 1979, The Usefulness of Optimal Control Theory to Ecological Problems, in: Theoretical Systems Ecology, E. Halfon, ed., Academic Press, N.Y., 385-399.

Hannon, B. 1973, The Structure of Ecosystems, J. Theo. Biol., **41**, 535-46.

Hannon, B 1979, Total Energy Costs in Ecosystems, J. Theo Biol., **80**, 271-293.

Hannon, B. 1985a, Ecosystem Flow Analysis, Canadian Journal of Fisheries and Aquatic Sciences 213, Ecological Theory for Biological Oceanography, R. Ulanowicz and T. Platt, eds., 97-118.

Hannon, B. 1985b, Conditioning the Ecosystem, Mathematical Biosciences, 75, 1, 23-42.

Hannon, B. 1985c, Linear Dynamic Ecosystems, J. Theo. Biol., **116**, 1, 89-98.

Hannon, B. 1986, Ecosystem Control Theory, J. Theo. Biol., **121**, 417–437.

Herendeen, R. 1981, Energy Intensities In Ecological and Economic Systems, J. Theo. Biol. **91**, 607-620.

Kemp, W. and W. Mitsch, 1979, Turbulence and Phytoplankton Diversity: A General Model of the "Paradox of Plankton", Ecol. Model., **7**, 201–222.

Luenberger, D. 1979, Introduction to Dynamic Systems: Theory, Models, and Application, Wiley, NY.

Levine,S. 1977, Exploitation Interactions and the Structure of Ecosystems, J. Theo. Biol., **69**, 345-355.

Levine, S. 1980, Several Measures of Trophic Structure Applicable to Complex Food Food Webs, J. Theo. Biol., **83**, 195-207.

Levines, R., 1979, Coexistence in a Variable Environment, Amer. Natur., **11<**, 6, 765–783.

Lowes, A. and C. Blackwell, 1975, Applications of Modern Control Theory to Ecological Systems, in: Ecosystems Analysis and Prediction, S. Levin, ed., SIAM, Phil., Pa., 299-305.

Meerkov, S. M., 1980, Principle of Vibrational Control: Theory and Applications, IEEE Trans. Auto. Control, **AC-25**, 4, 755-762.

Mulholland, R. and C. Sims, 1976, Control Theory and Regulation of Ecosystems, in: Systems Analysis and Simulation in Ecology, Patten, B., ed., Academic Press, NY., 373-390.

Olsen, J. 1961, Analog Computer Models for Movement of Nuclides Through Ecosystems, Radioecology, Proc. First National Symp., Colo. State University, Schultz and Klement, A. eds., Rheinhold Publishing Co., N.Y., 121-125

Patten, B., R. Bosserman, J. Finn, and W. Cale, 1976, Systems Analysis and Simulation in Ecology, 4, 457-574.

Takahashi, Y., M. Rabins and D. Auslander, 1970, Control and Dynamic Systems, Addison–Wesley Publishing, Reading, MA, 348.

Ulanowicz, R. 1984, Growth and Development: A Phenomenological Perspective, Center for Environmental And Estuarine Studies, Chesapeake Biological Laboratory, University of Maryland, Solomons, MD.

Vincent, T.L., C.S. Lee and B.S. Goh, 1977, Control Targets for the Management of Biological Systems, Ecological Modeling, 3, 285–300.

Wonham, W., 1984, Regulation, Feedback and Internal Models, in: Adaptive Control of Ill–Defined Systems, O. Selfridge and E. Rissland, eds., Plenum Press, NY., 75–88.

Wu, M., 1975, Some Results in Linear Time Varying Systems, IEEE AC–20, 159–161.

SELF CONTROLLED GROWTH POLICY
FOR A FOOD CHAIN SYSTEM

George Bojadziev
Department of Mathematics and Statistics
Simon Fraser University
Burnaby, B.C. V5A 1S6
Canada

Abstract. A behavioural policy of controlled growth for a food chain model of length 2n is considered. The highest trophic level population controls its own growth in order to restrain the growth of the other 2n-1 populations in the system so as to avoid undesirable outcomes.

1. INTRODUCTION

The present research concerning control policies for biological systems in population dyna.:.cs mainly deals with human control added to models of interacting populations. Various pest management programs provide typical examples of this kind of *external* control [1,2]. However in reality there are also situations in which one or more populations participatirg in the system are the controllers. Such systems change behaviour abruptly in response to changes of the size of the interacting populations, climatic conditions, diseases, etc. We call this type of control *internal*. The classical models in population dynamics usually do not reflect either the external nor the internal control. The controlling populations can apply the internal control to their own members (*self control*) or to all or some of the other participating populations in the model. In this paper the attention is focused on the concept of self control.

Generalizing a previous paper (Bojadziev and Skowronski [3]) here we study a food chain system of size 2n involving a controlling factor u(t) which adjusts the number of the highest trophic level population so that a reasonable size of all populations is maintained. Making use of a methodology developed by Leitmann and Skowronski [4] (see also Blaquière, Gerard, and Leitmann [5]) for dynamical systems, we derive conditions under which the designed control policy results in avoidance of a prescribed region in R^{2n} so that undesirable outcomes are avoided.

2. THE FOOD CHAIN MODEL

Consider the food chain model with control

$$\bar{x}'(t) = f(\bar{x}(t),\ u(t)) \tag{1}$$

where $t \in R_+$ is the time variable, $\bar{x}(t) = (x_1,\dots,x_{2n})^T$ is the population vector, $u(t)$ is the control, and the components of the vector function $\bar{f}(x,u) = (f_1,\dots,f_{2n})^T$ are given by

$$\begin{aligned}
f_1(\bar{x},u) &= x_1(\alpha_1 - \frac{\beta_1}{\gamma_1}x_2)\ , \\
f_{2k}(\bar{x},u) &= x_{2k}\left(-\ \alpha_{2k} + \frac{\beta_{2k-1}}{\gamma_{2k}}x_{2k-1} - \frac{\beta_{2k}}{\gamma_{2k}}x_{2k+1}\right), \\
f_{2k+1}(\bar{x},u) &= x_{2k+1}\left(-\ \alpha_{2k+1} + \frac{\beta_{2k}}{\gamma_{2k+1}}x_{2k} - \frac{\beta_{2k+1}}{\gamma_{2k+1}}x_{2k+2}\right), \\
f_{2n}(\bar{x},u) &= x_{2n}\left(-\ \alpha_{2n} + \frac{\beta_{2n-1}}{\gamma_{2n}}x_{2n-1}\right) + ux_{2n}^2\ , \\
k=1,\dots,n-1,\quad & f_i(\bar{x},u) = f_i(\bar{x},0),\quad i=1,\dots,2n-1\ .
\end{aligned} \tag{2}$$

For $u=0$ the model (1) reduces to the uncontrolled food chain model

$$\bar{x}'(t) = \bar{f}(\bar{x}(t),0)\ . \tag{3}$$

In (1) $x_i, i=1,\dots,2n$, is the size of the i-th population; α_i (growth rate coefficient), β_i (interaction coefficient), and γ_i (trophic weight factor) are positive constants; γ_j/γ_i expresses the gain-loss ratio when population i interacts with population j. The control $u(t) \in \mathcal{U}[t_o,\hat{t}] = \{u(t): u(t) \in U$ and $u(t)$ measurable on $[t_o,\hat{t}]\}$, $0 \le t_o < \hat{t} < \infty$, $U \subset R$ is a compact set to be specified later in accordance to a growth restriction policy.

The biological meaning of the control term ux_{2n}^2 in the last expression (2) which takes part in (1) is that for $u > 0$ the population with size x_{2n} (the highest trophic level population in the food chain) is enhanced by increasing the population density (increasing returns) and for $u < 0$ it dampers its own growth (diminishing returns). The 2n-th population can be considered as a consumer or predator of a higher level in terms of organization and brain capability in comparison to the other 2n-1 populations or resources. The self controlled growth of the consumer (predator) will affect the growth of all populations in the food chain system.

Each choice of control, say $u(t_o) = c_o \in U$ on some time interval starting at $t = t_o$, generates a solution or response $k[t] = k(\bar{x}(t_o), c_o, t)$ of the system (1) with initial state $\bar{x}(t_o) \in R^{2n}_+$ which geometrically is represented by an orbit ℓ_o in the phase space R^{2n}. If $c_o = 0$ (no control, hence (1) reduces to (3)) the response $k(x(t_o), 0, t)$ of (3) can exhibit large variation and may endanger the existence of an acceptable size of some populations. In order to avoid such undesirable outcomes, the consumer population with size x_{2n} may opt to self control its own growth which will affect the growth of the other populations in the food chain. This can be accomplished by selecting a suitable control value $u(t_1) = c_1 \in U$ at a point $\bar{x}(t_1) \in R^{2n}_+$ (switching point) on some time inverval starting at $t = t_1$, $t_1 > t_o$. The control value $u(t_1) = c_1$ will generate a response $k(\bar{x}(t_1), c_1, t)$ along a new orbit ℓ_1, $\ell_0 \cap \ell_1 = \bar{x}(t_1)$.

Using a Liapunov function for the uncontrolled model (3) we define for the response of (1) an avoidance region A, a security zone S which safeguards the response of entering A, and design a control policy for avoidance.

3. THE LIAPUNOV FUNCTION

The coordinates of the nontrivial equilibrium $E^0(\bar{x}^0)$, $\bar{x}^0 = (x_1^0, \ldots, x_{2n}^0)^T \in R^{2n}$, of (1) are

$$x_2^0 = \frac{\alpha_1\gamma_1}{\beta_1}, \quad x_{2n-1}^0 = \frac{\alpha_{2n}\gamma_{2n}}{\beta_{2n-1}},$$

$$x_{2k-1}^0 = \frac{\alpha_{2k}\gamma_{2k} + \beta_{2k}x_{2k+1}^0}{\beta_{2k-1}}, \quad k=1,\ldots,n-2, \tag{4}$$

$$x_{2k+2}^0 = \frac{-\alpha_{2k+1}\gamma_{2k+1} + \beta_{2k}x_{2k}^0}{\beta_{2k+1}}, \quad k=1,\ldots,n-1.$$

We require that $E^0 \in \text{Int } R^{2n}_+$, the interior of the closed positive cone, so that E^o has biological meaning. Since $x_{2n-1}^0 > 0$, it follows from (4) that $x_{2k-1}^0 > 0$, $k-1,\ldots,n-1$. Also from (4) we·see that $x_2^0 > 0$. However, in order to secure that $x_{2k}^0 > 0$, $k=2,\ldots,n-1$, we assume that $x_{2k}^0 > \alpha_{2k+1}\gamma_{2k+1}/\beta_{2k}$.

The model (3) has the Volterra function (Huang and Morowitz [6])

$$V(\bar{x}) = \sum_{i=1}^{2n} \gamma_i x_i^0 \left(\frac{x_i}{x_i^0} - \ell n \frac{x_i}{x_i^0} - 1 \right), \tag{5}$$

continuous on Int R_+^{2n}, which is actually a Liapunov function with the following properties.

(i) The minimum of $V(\bar{x})$ is attained at the equilibrium $E^0(\bar{x}^0)$ given by (4); $\min V(\bar{x}^0) = 0$;

(ii) $V(\bar{x})$ is monotone increasing about E^0 (has the nesting property);

(iii) $$\frac{dV(\bar{x})}{dt} = \sum_{i=1}^{2n} \frac{\partial V}{\partial x_i} f_i(\bar{x},0) = 0 \ , \tag{6}$$

where f_i are given by (2). From here follows that the equilibrium $E^0(\bar{x}^0)$ is stable.

The model (3) has a first integral

$$V(\bar{x}) = h, \qquad h = \text{const} > 0 \ , \tag{7}$$

which represents a family of level surfaces V_h in R^{2n+1}. The orthogonal projection of V_h onto R^{2n} generates 2n dimensional hypersurfaces H_h in R^{2n} which are closed, do not intersect, contain inside the equilibrium E^0, and accommodate orbits of (1). Further, if $h_1 < h_2$, the hypersurface H_{h_1} is inside the hypersurface H_{h_2}.

4. AVOIDANCE CONTROL

Here, marking use of a Liapunov design technique [4], we introduce definitions and prove a theorem concerning the food chain model (1).

Definition 1 (Avoidance set A). Given $\bar{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_{2n})^T \in \text{Int } R_+^{2n}$ and the Liapunov function $V(\bar{x})$ by (5),

$$A \triangleq \{\bar{x} \in R^{2n}: V(\bar{x}) \geq V(\bar{\varepsilon}) = h_\varepsilon\} \ , \tag{8}$$

where ε_i (avoidance parameters), i=1,...,2n, are small as desired for a particular study. The boundary of A is

$$\partial A = H_{h_\varepsilon} \triangleq \{\bar{x} \in R^{2n}: V(\bar{x}) = h_\varepsilon\} \ . \tag{9}$$

Definition 2 (Security zone S). Given $\bar{\delta} = (\delta_1,\ldots,\delta_{2n})^T \in \text{Int } R^{2n}$, $\delta_i > \varepsilon_i$, and $V(\bar{x})$ by (5),

$$S \triangleq \{\bar{x} \in R^{2n}: V(\bar{x}) \geq V(\bar{\delta}) = h_\delta\} - A \ , \tag{10}$$

δ_i, i=1,...,2n, are security parameters. The boundary of S is given by

$$\partial S = H_{h_\delta} \triangleq \{\bar{x} \in R^{2n}: V(\bar{x}) = h_\delta\} \ . \tag{11}$$

From the nesting property of $V(\bar{x})$ it follows that $h_\delta < h_\varepsilon$, hence in

R_+^{2n} the hypersurface (9) encloses the hypersurface (11).

Definition 3 The set A defined by (8) is avoidable if there is a set S defined by (10) and a control $u \in U$ such that for all $\bar{x}^s(t_s) \in S$, the response $k(\bar{x}^s(t_s), u(t_s), t)$ of (1) cannot enter A, i.e.

$$k(\bar{x}^s(t_s),\ u(t_s),\ t) \cap A = \phi \ \forall\ t\ . \tag{12}$$

Now we establish sufficient conditions for the avoidance of A.

Theorem The food chain model (1) is controllable for avoidance of A if there is a control $u(t) \in U$ and a Liapunov function $V(\bar{x})$ defined by (5) so that

$$\frac{dV(\bar{x})}{dt} = \sum_{i=1}^{2n} \frac{\partial V}{\partial x_i} f_i(\bar{x},u) \le 0\ , \tag{13}$$

where $f_i(\bar{x},u)$ are given by (2).

Proof. Assume that A is not avoidable, i.e. (12) is violated. Hence for some $\bar{x}^s(t_s) \in S$, the response $k(\bar{x}^s(t_s),\ u(t_s), t)$ enters A, $t > t_s$. Then there is a $t_a > t_s$ for which $\bar{x}^a(t_a) = k(\bar{x}^s(t_s), u(t_s), t_a) \in \partial A$. From the nesting property of $V(\bar{x})$ it follows that $V(\bar{x}^s(t_s)) < V(\bar{x}^a(t_a))$, meaning that the function $V(\bar{x})$ is increasing. This contradicts (13) which states that $V(\bar{x})$ is non-increasing along every response of (1).

5. THE CONTROL POLICY

To design a policy for avoidance the region A by the response of (1) we use the theorem in the previous section. Substituting $f_i(\bar{x},u)$ from (2) into (13) with (5) gives

$$\frac{dV(\bar{x})}{dt} = \sum_{i=1}^{2n} \frac{\partial V}{\partial x_i} f_i(\bar{x},0) + \frac{\partial V}{\partial x_{2n}} u x_{2n}^2 \le 0\ .$$

According to (6) the summation term above is zero; the second term gives

$$\gamma_{2n} x_{2n}^0 \left(\frac{1}{x_{2n}^0} - \frac{1}{x_{2n}} \right) u x_{2n}^2 \le 0$$

which can be written as

$$\left(\frac{1}{x_{2n}^0} - \frac{1}{x_{2n}} \right) u \le 0\ . \tag{14}$$

The inequality (14) establishes a relationship between the control u and the controlling population x_{2n}. It requires that

$$u \leq 0 \ \forall \ x_{2n} > x_{2n}^0 ,$$
$$u \geq 0 \ \forall \ x_{2n} < x_{2n}^0 . \qquad (15)$$

According to (15) we specify that

$$u(t) \in U = [-r,r] \subset R, \quad r=\text{const.} \qquad (16)$$

On the basis of (15) we formulate the following behavioural policy.

Avoidance control policy: If the response $k[t] = k(\bar{x}(t_o),u(t_o),t)$ of the food chain model (1) with initial state $\bar{x}(t_o)$ and fixed control $u(t_o) \in U$, U specified by (16), enters the security zone S given by (10), in order to prevent $k[t]$ of entering into A defined by (8), a new control value $u(t_s)$ should be selected from U at a switching point $\bar{x}(t_s) \in S$ with corresponding response $k(\bar{x}(t_s),u(t_s),t)$, $t_s > t_o$. If $x_{2n} > x_{2n}^0$, the new control value $u(t_s)$ should be negative and if $x_{2n} < x_{2n}^0$, it should be positive.

Note 1. The control $u=0$ satisfies (15) but then the response will be accommodated on a hypersurface H_{h_s} enclosed in the security zone S, $H_{h_\zeta} < H_{h_s} < H_{h_\varepsilon}$, which may not be satisfactory since large population fluctuations occur.

Note 2. The particular situation $x_{2n} = x_{2n}^0$ at $\bar{x}(t_s) \in S$ satisfies (14), hence any value $u \in U$ can be selected temporarily until the response moves to a neighbouring point in S for which $x_{2n} \neq x_{2n}^0$. Then the avoidance control policy can be applied.

LITERATURE REFERENCES

1. Vincent, T.L. Pest management programs via optimal control theory. Biometrics, 31, 1-10 (1975).
2. Goh, B.S., G. Leitmann, and T.L. Vincent. Optimal control of a prey-predator system. Math. Biosc., 19, 263-286 (1974).
3. Bojadziev, G. and J. Skowronski. Controlled food consumption. Methods of Operations Research, 49, 499-506 (1985).
4. Leitmann, G. and J. Skowronski. Avoidance control. J. Optim. Theory and Appl., 23, 581-591 (1977).
5. Blaquiere, A., F. Gerard and C. Leitmann. Quantitative and Qualitative Games. Academic Press, New York, 1969.
6. Huang, H.-W. and H. Morowitz. A method for phenomenological analysis of ecological data. J. theor. Biol., 35, 489-503 (1972).

MATHEMATIQUES ET SYSTEMES, ASPECT CALCUL

MATHEMATICS AND SYSTEMS, COMPUTATIONAL BEARINGS

QUASILINEARIZATION IN BIOLOGICAL SYSTEMS MODELING

E. S. Lee* and K. M. Wang**

The estimation of parameters in differential equations is a basic problem in biological systems modeling. However, these parameters cannot be estimated easily when the equations are too complicated and cannot be solved in closed form. Although Dr. Bellman has proposed to use quasilinearization to solve this problem, more numerical experiments are needed to show the effectiveness of this approach. In this paper, quasilinearization is used to estimate the parameters in various biological models. It is shown that this approach is quite effective and converges very fast in most situations. Thus, the quadratic convergence property is preserved.

QUASILINEARIZATION AND THE NONLINEAR ESTIMATION OF PARAMETERS

The algorithm of quasilinearization in estimation is well documented [1-3], only the essential equations will be discussed in the following. Consider a system represented by the following system of nonlinear differential equations

$$\frac{dx}{dt} = f(x, \alpha, t) \tag{1}$$

where x and f are M-dimensional vectors with components $x_1, x_2, \ldots, x_M$ and $f_1, f_2, \ldots, f_M$, respectively and α represents the L dimensional unknown parameters. Let us assume that the L parameters cannot be measured directly and only M_1 of the M variables can be measured. These measured values are

$$x_j^{(exp)}(t_s) = b_s^{(j)}, \quad S = 1,2,\ldots,m, \quad j = 1,2,\ldots,M_1 \tag{2}$$

with $t_m = t_f$. The problem is to estimate the parameters $\alpha_\ell(t)$, $\ell = 1,2,\ldots,L$ and the initial conditions

$$x_i(0) = c_i, \quad i = 1,2,\ldots,M \tag{3}$$

from the given or measured data, Equation (2). It should be emphasized that the measured values $b_s^{(j)}$ do contain noise. Let us establish the vector equation

$$\frac{d\alpha}{dt} = 0 \tag{4}$$

* Corresponding author, E. S. Lee, Dept. of Ind. Engg., Kansas State University, Manhattan, KS 66506
** Dept. of Ind. Engg., Tsinghua University, Taiwan, China

The problem can be stated as find the values of the vectors c and α so that the least square expression

$$J = \sum_{j=1}^{M_1} \sum_{s=1}^{m} \left\{x_j(t_s) - b_s^{(j)}\right\}^2 \tag{5}$$

is minimized subject to the constraints of Equations (1) and (4). This is a multipoint boundary value problem with minimization. It can be solved by the use of quasilinearization . Equations (1) and (4) can be combined to obtain

$$\frac{dy}{dt} = g(y,t) \tag{6}$$

where y and g are M + L dimensional vectors. Equation (6) can be linearized by the use of Taylor Series with second and higher order terms omitted. The resulting vector equation is

$$\frac{dy_{k+1}}{dt} = g(y_k, t) + J(y_k)(y_{k+1} - y_k) \tag{7}$$

where y_k is assumed known and is obtained from the previous iteration and Y_{k+1} is the unknown function. The expression $J(y_k)$ is the Jacobian matrix. Because of the fast convergence rate, Equation (7) with unknown initial conditions can be solved quickly by the use of the superposition principle. In general, less than ten iterations are needed to obtain a very high accuracy.

THE ARTIFICIAL KIDNEY SYSTEM

Consider the following simple model of the artificial kidney system [4, 10].

$$V_1 \frac{dC_1}{dt} = G - K(C_1 - C_2) \tag{8}$$

$$V_2 \frac{dC_2}{dt} = K(C_1 - C_2) - C_k C_2 - C_d C_2 \tag{9}$$

where G = urea (or creatinine) production rate

k = mass transfer parameter

C_k = clearance rate of patient kidney

C_d = dialyzer clearance

C_1 = urea concentration in intracellular cell

C_2 = urea concentration in extracellular cell

V_1 = volume of intracellular cell

V_2 = volume ot extracellular cell

In actual experimental situations, the constants or parameters cannot be measured, only C_2 can be measured at the various values of t. Our problem is to estimate k and $C_1(0)$ for Equations (8) and (9) from the experimental data

$$C_2^{(exp)}(t_s) = C_{2s}, \qquad s = 1,2,\ldots,m \tag{10}$$

Notice that the initial condition of $C_2(t=0)$ can be measured, but $C_1(t=0)$ must be estimated. Thus, an equation like Equation (4) can be established for the parameter k.

This problem is solved by quasilinearization with the following experimental data [4]

$$C_2^{(exp)}(t_s=1) = 2.070,$$

$$C_2^{(exp)}(t_s=2) = 1.818$$

$$C_2^{(exp)}(t_s=3) = 1.674$$

and the values of

$$G = 0.031, \quad C_d = 3.6, \quad C_k = 0, \quad \Delta t = 0.01,$$

$$C_2(t=0) = 2.538, \; t_f = 3$$

Four different experiments were carried out with four different sets of initial approximations. The convergence rates are summarized in Table 1. Notice that five digits accuracy are obtained in 6 to 10 iterations. The Runge-Kutta integration technique was used.

GLUCOSE AND INSULIN KINETICS MODELING

Consider the following simple one compartment model of glucose and insulin in plasma [5, 6]

$$\frac{dH}{dt} = - I_1H + I_3G + I_2 \tag{11}$$

$$\frac{dG}{dt} = - I_4G - I_6H + I_5 \tag{12}$$

where G = plasma glucose concentration

H = plasma IRI concentration

I_i = parameters or constants.

The problem is to estimate I_1, I_3, I_4, I_6, $H(t=0)$ and $G(t=0)$ from experimental data for H and G at various values of t. Again, equations like equation (4) can be established for the four parameters.

The four parameter values and the two initial conditions are estimated by quasilinearization. The numerical values used are

$I_2 = -1.56, \quad I_5 = 6.94, \quad t_f = 180$ minutes

$\Delta t = 0.2.$

The experimental data used are listed in Table 2. Several different sets of initial approximations are used. One of the typical results are listed in Table 3. The initial approximations are obtained by integrating the equations with the values for the Zeroth iteration as the initial conditions. The Runge-Kutta technique is again used. Notice that even with the very extreme initially assumed initial conditions of zero, only nine iterations are needed to obtain a five digits accuracy.

CARDIOVASCULAR INDICATOR DILUTION MODELING

Consider the following four cell cardiovascular indicator dilution model [7, 8].

$$\frac{dC_1}{dt} = B_1C_1 + B_2C_4$$

$$\frac{dC_2}{dt} = B_1(C_1 - C_2)$$

$$\frac{dC_3}{dt} = B_1(C_2 - C_3) \qquad (13)$$

$$\frac{dC_4}{dt} = B_1(C_3 - C_4)$$

where $B_1 = F/V, \quad B_2 = F_s/V$, with

F = volumetric flow rate

F_s = recycle volumetric flow rate

V = volume of the well-mixed cells

The boundary conditions for Equation (13) are

$$C_1(t=0) = \frac{M}{V} = B_3, \qquad C_2(t=0) = 0$$

$$C_3(t=0) = 0 \qquad C_4(t=0) = 0 \qquad (14)$$

where M is the mass of the injection and the C_i's are the concentrations of the corresponding cells.

In actual experiments, only the C's can be measured, the parameters B_1 and B_2 cannot be measured directly and must be estimated indirectly from experimental data.

The values of B_1, B_2 and B_3 are estimated by quasilinearization with the numerical data listed in Table 4. The Runge-Kutta numerical integration formula with $\Delta t = 0.2$ is used. Various different initial approximations for B_1, B_2 and B_3 were used. The convergence rate is again very fast. Three typical convergence results are listed in Table 5 for three different sets of initial approximations.

METHOTREXATE PHARMACOKINETICS MODELING

Consider the following pharmacokinetic model used to predict the detailed distribution and excretion of methotrexate in mammalian species over a wide range of doses [9]. The material balance equations representing the various anatomical compartments are

$$\text{Plasma:} \quad V_p \frac{dC_p}{dt} = Q_L \frac{C_L}{R_L} + Q_K \frac{C_K}{R_K} + Q_M \frac{C_M}{R_M} - (Q_L + Q_K + Q_M)\ C_p \tag{15}$$

$$\text{Muscle:} \quad V_M \frac{dC_M}{dt} = Q_M \left(C_p - \frac{C_M}{R_M}\right) \tag{16}$$

$$\text{Kidney:} \quad V_K \frac{dC_K}{dt} = Q_K \left(C_p - \frac{C_K}{R_K}\right) - k_K \frac{C_K}{R_K} \tag{17}$$

$$\text{Liver:} \quad V_L \frac{dC_L}{dt} = (Q_L - Q_G)\left(C_p - \frac{C_L}{R_L}\right) + Q_G \left(\frac{C_G}{R_G} - \frac{C_L}{R_L}\right) - r \tag{18}$$

$$\text{Gut Tissue:} \quad V_G \frac{dC_G}{dt} = Q_G \left(C_p - \frac{C_G}{R_G}\right) + 1/4 \sum_{i=1}^{4} \left(\frac{k_G \quad C_i}{K_G + C_i} + b\ C_i\right) \tag{19}$$

$$\text{Gut Lumen:} \quad \frac{dC_{GL}}{dt} = 1/4 \sum_{i=1}^{4} \frac{dC_i}{dt} \tag{20}$$

$$\frac{V_{GL}}{4} \quad \frac{dC_1}{dt} = r_3 - k_F \quad V_{GL} \quad C_1 - 1/4\left(\frac{k_G \quad C_1}{K_G + C_1} + b\ C_1\right) \tag{21}$$

$$\frac{V_{GL}}{4} \quad \frac{dC_i}{dt} = k_F \quad V_{GL} \ (C_{i-1} - C_i) - 1/4\left(\frac{K_G \quad C_i}{K_G + C_i} + b\ C_i\right) \tag{22}$$

$$i = 2,3,4$$

where the value of r in Equation (18) can be represented by

$$r = \frac{K_L(C_1/R_L)}{K_L+(C_L/R_L)} \tag{23}$$

which is the secretion rate of methotrexate out of the liver cells into the bile ducts. Using the three compartments model, we have

$$\tau \quad \frac{dr_1}{dt} = r - r_1 \tag{24}$$

$$\tau \quad \frac{dr_2}{dt} = r_1 - r_2 \tag{25}$$

$$\tau \quad \frac{dr_3}{dt} = r_2 - r_3 \tag{26}$$

where C is the drug concentration in the various anatomical compartments, r is the drug transport rate in the bile, V is the volume of the various compartments, b is the rate constant for nonsaturable gut absorption, Q is the plasma flow rate, R is the tissue plasma equilibrium ratio for linear binding and K_k is kidney clearance and is equal to 1.1 mℓ/min for rat. The other numerical values used for rat are:

$V_p = 9$ mℓ	$Q_k = 5$ mℓ/min
$V_M = 100$ mℓ	$Q_L = 6.5$ mℓ/min
$V_k = 1.9$ mℓ	$Q_G = 5.3$ mℓ/min
$V_L = 8.3$ mℓ	$R_M = 0.15$
$V_G = 11$ mℓ	$R_k = 3.0$
$V_{GL} = 11$ mℓ	$R_L = 3.0$
$Q_M = 3$ mℓ/min	

The body weight for rat is 200 g. Notice that three compartments were assumed for bile secretion and 4 compartments were assumed for gut lumen. Some of the parameters such as R_G, k_G and K_G are not measurable. These parameters for methotrexate in rat will be estimated by quasilinearization using experimental data obtained by Bischoff et al. [9]. These experimental data as a function of time for the drug concentrations in the various compartments are listed in Table 6 and are obtained from the figures of reference [9].

It should be emphasized that the parameters R_g, k_G and K_G cannot be estimated easily. This is because that the systems of differential equations cannot be solved in closed form. thus, quasilinearization forms an ideal and powerful approach.

In addition to the 13 differential equations represented by Equations (15) - (26), 3 additional differential equations in the form of Equation (4) can be formulated for the 3 unknown parameters. Thus, there are a total of 16 differential equations. The initial conditions for the 13 differential equations are all equal to zero except $C_p(t)$ which is

$$C_p(t) = 1200/9 \tag{27}$$

The 16 different equations can be linearized by using Equation (7). The unknown parameters can then be obtained by using Equation (5) and superpositoin principle. The homogeneous and particular solutions cau be obtained by numerically integrating the linearized equations. In the present work, the modified Adam-Moulton integration scheme is used with step size as

$$\Delta t = 0.01 \text{ minute for } 0 \le t \le 30$$

$$\Delta t = 0.1 \text{ minute for } 30 \le t \le 240.$$

The convergence rates for the three parameters are listed in Table 7. Notice the fast convergence rates. Only 5 iterations are needed to obtain 4 digits accuracy.

DISCUSSION

Since the results of the previous iteration for all t must be stored in the computer, the storage requirement can be quite large. For example, the pharmacokinetic model needs (30/0.01 + 210/0.1 + 1) 16 = 81616 storage spaces. In order to reduce this storage requirement, we can store only the initial conditions of the previous iteration. The complete profile for all t of the previous iteration can be obtained by integrating the equations when we calculate the current iterations. The storage requirements can thus be reduced tremendously. For the pharmacokinetic problem, the storage requirement is reduced from 81616 to 16.

REFERENCES

[1] Bellman, R. and R. Kalaba, Quasilinearization and nonlinear Boundary Value Problems, American Elsevier, NY (1965).

[2] Lee, E. S., Quasilinearization and Invariant Imbedding, Academic Press, NY (1968).

[3] Lee, E. S., "Quasilinearization" The Bellman Continuum, World Scientific Publishing Co. (1986).

[4] Bell, R. L., F. K. Curtis and A. L. Babb, "Analog Simulation of the Patient-Artificial Kidney System" Trans. Am. Soc. Artificial Internal Organs, 11, 183 (1965).

[5] Ackerman, E., L. C. Gatewood, J. W. Rosevear and G. D. Molnar, "Model Studies of Blood Glucose Regulation," Bull. Math. Biophys, 27, 21 (1965).

[6] Norwich, K. H. "Mathematical Models of the Kinetics of Glucose and Insulin in Plasma," Bull. Math. Biophys., 31, 105 (1969).

[7] Harris, T. R., "The identification of recirculating systems in the frequency domain," Bull. Math. Biophys., 30, 87 (1968).

[8] Nicholes, K. K. and H. R. Warner, "Study of dispersion of an indicator in the circulation," Ann. N.Y. Acad. Sci., 115, 721 (1964).

[9] Bischoff, K. B., R. L. Dedrick, D. S. Zaharko and J. A. Longstreth, "Methotrexate Pharmacokinetics," J. Pharmaceutical Science, 60, 1128 (1971).

[10] Abbrecht, P. H. and N. W. Prodany "A Model of the Patient-Artificial Kidney System," IEEE Trans. Bio-Medical Eng., BME-18, 257 (1971).

Table 1 Convergence Rates of the Artificial Kidney Model

Iteration	$C_1(0)$	K	$C_1(0)$	K	$C_1(0)$	K	$C_1(0)$	K
0	2.538	5.	2.538	12.	2.538	19.2	2.538	25.
1	2.9513	6.1718	2.7879	5.2057	3.1695	-35.947	2.4352	18.639
2	2.7675	7.5204	2.8314	7.4735	2.9149	- 4.7906	3.1274	-34.37
3	2.7997	7.5318	2.8023	7.4970	2.9895	6.9627	2.7892	- 7.5045
4	2.8000	7.5279	2.7994	7.5351	2.7776	7.5923	2.6165	5.6438
5	2.7999	7.5288	2.8000	7.5272	2.7991	7.5369	2.8398	7.8420
6	2.7999	7.5286	2.7999	7.5289	2.8000	7.5270	2.8016	7.5273
7	2.7999	7.5286	2.7999	7.5285	2.7999	7.5290	2.7999	7.5292
8			2.7999	7.5286	2.7999	7.5285	2.7999	7.5285
9			2.7999	7.5286	2.7999	7.5286	2.7999	7.5286
10					2.7999	7.5286	2.7999	7.5286

Table 2 Experimental Data for Glucose and Insulin Kinetics Model

t_s	$H^{(exp)}(ts)$	$G^{(exp)}(ts)$
0	177	581
30	155	182
60	40	95
90	26	87
120	20	97
150	24	106
180	28	110

Table 3 Convergence Rates of Glucose and Insulin Kinetics Model

Iteration	I_1	I_2	I_4	I_6	H(0)	G(0)
0	0.	0.	0.	0.	177.	581.
1	0.051076	0.025872	0.048153	0.22224	181.31	576.58
2	0.038405	0.017182	0.020605	0.052089	177.16	580.37
3	0.045445	0.021543	0.028009	0.043957	177.56	580.06
4	0.046151	0.022149	0.028790	0.043174	177.27	580.35
5	0.046411	0.022281	0.028581	0.043500	177.24	580.38
6	0.046408	0.022286	0.028565	0.043510	177.23	580.39
7	0.046423	0.022293	0.028555	0.043523	177.23	580.39
8	0.046421	0.022292	0.028555	0.043523		
9	0.046422	0.022293	0.028555	0.043524		
10	0.046422	0.022293	0.028555	0.043524		

Table 4 Experimental Data for Cardiovascular Model

t_s	$C_1(t_s)$	$C_2(t_s)$	$C_3(t_s)$	$C_4(t_s)$
0.0	0.9997	0.0	0.0	0.0
2.0	0.2289	0.3314	0.2609	0.1387
4.0	0.1327	0.1887	0.2391	0.2366
6.0	0.1141	0.1347	0.1682	0.2009
8.0	0.0909	0.1066	0.1269	0.1528
10.0	0.0702	0.0834	0.0988	0.1175
12.0	0.0543	0.0646	0.0768	0.0912
14.0	0.0421	0.0501	0.0595	0.0707
16.0	0.0327	0.0388	0.0462	0.0549
18.0	0.0253	0.0301	0.0358	0.0425
20.0	0.0196	0.0234	0.0278	0.0329

Table 5 Convergence Rate of Cardiovascular Model

Iter-ation	B1	B2	B3	B1	B2	B3	B1	B2	B3
0	0.1	0.01	0.1	0.6	0.2	0.8	2.	1.5	3
1	0.4379	0.0725	0.4969	0.7663	0.3755	0.9903	1.7167	1.2619	1.0049
2	0.4896	0.1772	0.8993	0.7966	0.3970	0.9992	0.6846	0.2886	0.9983
3	0.6522	0.2679	0.9658	0.8013	0.4015	0.9996	0.8014	0.4021	0.9992
4	0.7616	0.3635	0.9952	0.8017	0.4018	0.9997	0.8017	0.4018	0.9997
5	0.7974	0.3979	0.9993	0.8017	0.4018	0.9997	0.8017	0.4018	0.9997
6	0.8014	0.4015	0.9997						
7	0.8017	0.4018	0.9997						
8	0.8017	0.4018	0.9997						

Table 6 Experimental Data for Pharmacokinetics Modeling

t_s (min)	$C_P(t_s)$	$C_M(t_s)$	$C_K(t_s)$	$C_L(t_s)$	$C_{GL}(t_s)$
15	7.7	1.5	20.	20.9	23.98
30	4.0	0.75	10.8	11.5	47.00
60	1.5	0.25	4.0	4.97	59.00
90	1.14	0.16	2.8	3.60	45.50
120	0.80	0.13	2.2	2.80	36.00
180	0.45	0.072	1.1	1.45	18.25
240	0.27	0.043	0.67	0.86	8.90

Table 7 Convergence Rates of Pharmacokinetics Model

Iteration	R_G	k_G	K_G
0	1.	20.	200.
1	1.108	22.64	237.2
2	1.112	21.61	224.6
3	1.112	21.97	229.3
4	1.112	21.85	227.7
5	1.112	21.89	228.3
6	1.112	21.89	228.3

A THREE-MIRROR PROBLEM ON DYNAMIC PROGRAMMING

Seiichi Iwamoto

Department of Economic Engineering
Faculty of Economics
Kyushu University 27, Fukuoka 812, Japan

1. INTRODUCTION

The essence of dynamic programming states that a simultaneous optimization of real-valued two-variable functions is assured by the two-stage optimization under both separability and monotonicity [15, 16]. We call these two properties the recusiveness with monotonicity —— dynamic programming structure —— [8, 11]. This structure yields what we call dynamic programmable function [11].

In this paper we focus our attention on both dynamic programming structure and quasililearization for a class of objective functions. Given a differentiable strictly increasing convex function $f : R^1 \longrightarrow R^1$, we approximate $f(x)$ by its linear approximation $f(x;h)$ $R^1 \times R^1 \longrightarrow R^1$, which is strictly increasing in h for $x \in R^1$. Thus, $f(x)$ is a quasilinearization of $f(x;h)$. The N-times composition of $f(x_n;\cdot)$ generates a dynamic programmable function $F(x;h) : R^N \times R^1 \longrightarrow R^1$. Similarly, inverse function $f^{-1}(y)$, reverse function $f_{-1}(x;k)$ which is the inverse function of $f(x;h)$ with respect to h for fixed x, and conjugate function $f^*(y)$ also generate dynamic programmable functions $F^{-1}(y;k)$, $F_{-1}(x;k)$, and $F^*(y;h) : R^N \times R^1 \longrightarrow R^1$, respectively. Thus, the function f yields four —— main, inverse, reverse, and conjugate —— optimization problems on R^N. These problems are solved through dynamic programming approach. Some relations between them are established. Finally we illustrate two interesting examples from Bellman [1].

2. PROBLEMS

First of all let us consider the following famous problem [1, p. 102; 8, p.101; 10, p.18]:

$$\text{Max}\quad e^{x_1}(1-x_1)+e^{x_1+x_2}(1-x_2)+\ldots+e^{x_1+\ldots+x_N}(1-x_N)$$
$$+\,e^{x_1+\ldots+x_N}\times h$$

$$\text{s.t.}\quad -\infty < x_n < \infty \qquad 1 \leq n \leq N$$

where h is a real constant. We remark that the N-times iteration of

$$f(x;h) = e^x(1-x+h)$$

yields the objective function

$$f(x_1;f(x_2;\ldots;f(x_N;h)\ldots))$$
$$= e^{x_1}(1-x_1)+e^{x_1}\left[e^{x_2}(1-x_2)+e^{x_2}\left[\ldots+e^{x_{N-1}}\right.\right.$$
$$\left.\left.\times\left[e^{x_N}(1-x_N)+e^{x_N}\times h\right]\ldots\right]\right]$$

(See also [11, p.278; 12, p.285]).

Second we consider the following maximization problem:

$$\text{Max}\quad (1-2x_1^2)\exp(x_1^2)+2x_1(1-2x_2^2)\exp(x_1^2+x_2^2)+4x_1x_2$$
$$\times(1-2x_3^2)\exp(x_1^2+x_2^2+x_3^2)+8x_1x_2x_3(1-2x_3^2)$$
$$\times\exp(x_1^2+x_2^2+x_3^2)h$$

$$\text{s.t.}\quad x_1\geq 0,\ x_2\geq 0,\ x_3\geq 0$$

where h ≥ 0. The three-times iteration of

$$f(x;h) = (1 - 2x^2 + 2xh)\exp(x^2)$$

generates

$$\begin{aligned}&f(x_1;f(x_2;f(x_3;h)))\\ &= (1-2x_1^2)\exp(x_1^2) + 2x_1\exp(x_1^2)\,[(1-2x_2^2)\exp(x_2^2) + 2x_2\exp(x_2^2)\times\\ &\quad[(1-2x_3^2)\exp(x_3^2) + 2x_3\exp(x_3^2)h]].\end{aligned}$$

These two functions are called *recursive functions on R^N(resp. R^3_+) with strict increasingness* ([10, 11]). A function $F : R^N \times R^1 \longrightarrow R^1$ is called *dynamic programmable function on R^N* if it is expressed as follows

$$\begin{aligned}&F(x_1,x_2,\ldots,x_N;h)\\ &= f_1(x_1;f_2(x_1,x_2;\ldots;f_N(x_1,x_2,\ldots,x_N;h)\ldots))\end{aligned}$$

where $f_n: R^n \times R^1 \longrightarrow R^1$ and $f_n(x_1,x_2,\ldots,x_n;\ \cdot): R^1 \longrightarrow R^1$ is nondecreasing for $1 \leq n \leq N$, $(x_1,x_2,\ldots,x_n) \in R^n$. Therefore, any recursive function with strict increasingness is a dynamic programmable function. In the following we are mainly concerned with a class of recusive functions on $X(\subset R^N)$ with strict increasingness.

3. MAIN RESULT

First, we prepare the following fundamental lemma. Let X and Y be two nonempty sets. For each $x \in X$ let $Y(x)$ be a nonempty subset of Y. That is, $Y(\cdot) : X \longrightarrow 2^Y$ is a point-to-set-valued mapping, where 2^Y denotes the set of all nonempty subsets of Y. Let

$$G_r(Y) = \{(x,y) \mid y \in Y(x),\ x \in X\} \subset X \times Y$$

be the graph of the mapping $Y(\cdot)$. In the following it will be clear from the context whether a notation Y is considered the set or the mapping.

LEMMA 1 (Maximax Theorem [11; p.268]) Let $f : X \times R^1 \longrightarrow R^1$ be a function such that $f(x;\cdot) : R^1 \longrightarrow R^1$ is nondecreasing for $x \in X$. Let $g : G_r(Y) \longrightarrow R^1$ be a function. If $\operatorname*{Max}_{x \in X} f(x; \operatorname*{Max}_{y \in Y(x)} g(x,y))$ exists, then $\operatorname*{Max}_{(x,y) \in G_r(Y)} f(x; g(x,y))$ exists and both are equal:

$$\operatorname*{Max}_{x \in X} f(x; \operatorname*{Max}_{y \in Y(x)} g(x,y)) = \operatorname*{Max}_{(x,y) \in G_r(Y)} f(x; g(x,y)).$$

REMARK This equality remains valid even if the operator Max is replaced by the operator min under the same condition as stated above. Furthermore, as a special case we have

$$\operatorname*{Max}_{-\infty<x<\infty} f(x; \operatorname*{Max}_{-\infty<y<\infty} g(y)) = \operatorname*{Max}_{-\infty<x,y<\infty} f(x; g(y)).$$

In general we have for any differentiable convex function $f : R^1 \longrightarrow R^1$

$$f(h) = \operatorname*{Max}_{-\infty<x<\infty} f(x;h) \qquad (1)$$

where

$$\left.\begin{aligned} f(x;h) &= F(x) + f'(x)h \\ F(x) &= f(x) - xf'(x). \end{aligned}\right. \qquad (2)$$

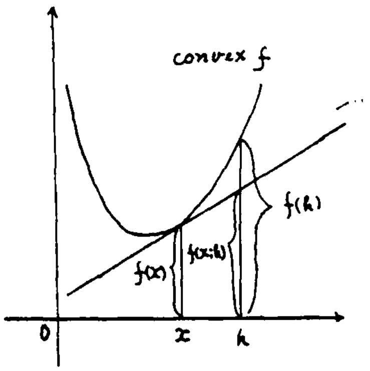


Fig. 1

Thus, $f(x;h)$ is the *linear approximation of* $f(x)$ *at* h:

$$f(x;h) = f(x) + (h - x)f'(x). \qquad (3)$$

The expression (1) is called a *quasilinearization of f(x)* ([1; p.135 ; 13; 14]).

Furthermore, from Lemma 1, we obtain under $f'(x) \geq 0, -\infty < x < \infty$

$$\begin{aligned} f(f(h)) &= \operatorname*{Max}_{-\infty<x_1<\infty} f(x_1; \operatorname*{Max}_{-\infty<x_2<\infty} f(x_2; h)) \\ &= \operatorname*{Max}_{-\infty<x_1,x_2<\infty} f(x_1; f(x_2;h)). \end{aligned} \qquad (4)$$

that is

$$\begin{aligned} f(f(h)) &= \operatorname*{Max}_{-\infty<x_1<\infty} \left[F(x_1) + f'(x_1)(\operatorname*{Max}_{-\infty<x_2<\infty} [F(x_2) + f'(x_2)h])\right] \\ &= \operatorname*{Max}_{-\infty<x_1,x_2<\infty} [F(x_1) + f'(x_1)F(x_2) + f'(x_1)f'(x_2)h]. \end{aligned} \qquad (5)$$

DEFINITION Let $f : R^1 \longrightarrow R^1$ be a differentiable increasing (resp. strictly increasing) convex function. Then we define $F : R^N \times R^1 \longrightarrow R^1$ by

$$\begin{aligned} F(x;h) &= f(x_1; f(x_2; \ldots; f(x_N;h)\ldots)) \\ &= F(x_1) + f'(x_1)F(x_2) + \ldots + f'(x_1)f'(x_2)\ldots f'(x_{N-1}) \\ &\quad \times F(x_N) + f'(x_1)f'(x_2)\ldots f'(x_N)h \end{aligned} \qquad (6)$$

where $f(x;h)$ and $F(x)$ are defined in (2), and $x = (x_1, x_2, \ldots, x_N)$. The function $F : R^N \times R^1 \longrightarrow R^1$ is the *recursive function with increasingness (resp. strict increasingness) generated by f* or simply *dynamic programmable function generated by f.*

In the following, it will be clear from the context a function f (resp. F) is considered f(x) or f(x;h) (resp. F(x) or F(x;h)).

REMARK The equalities (1) and (4) (or (5)) remain valid if we replace 'Max' and 'convex' with 'min' and 'concave', respectively. Similarly, a differentiable increasing (resp. strictly increasing) *concave* function $g : R^1 \longrightarrow R^1$ *generates the recursive function* $G : R^N{\times}R^1 \longrightarrow R^1$ *with increasingness (resp. strict increasingness)*, which is also called *dynamic programmable function generated by g:*

$$\begin{aligned} G(y;k) &= g(y_1;\ g(y_2;\ \ldots;\ g(y_N;k)\ldots)) \\ &= G(y_1) + g'(y_1)G(y_2) + \ldots + g'(y_1)g'(y_2)\ldots.g'(y_{N-1}) \\ &\quad \times G(y_N) + g'(y_1)g'(y_2)\ldots g'(y_N)k \qquad (7) \end{aligned}$$

where

$$y = (y_1,y_2,\ldots,y_N),$$

$$\begin{aligned} g(y;k) &= g(y) + (k - y)g'(y) \\ &= G(y) + g'(y)k, \end{aligned} \qquad (8)$$

$$G(y) = g(y) - yg'(y).$$

Therefore we have the following main result:

THEOREM 1. (i) Let $f: R^1 \longrightarrow R^1$ be a differentiable increasing convex function. Then for $h \in R^1$

$$f^N(h) = \operatorname*{Max}_{x \varepsilon R^N} F(x;h) \tag{9}$$

and $x_1^* = f^{N-1}(h)$, $x_2^* = f^{N-2}(h)$, ..., $x_{N-1}^* = f(h)$, $x_N^* = h$ attains the maximum, here and in the following $f^n(h)$ is the n-times composition of $f(x)$:

$$f^n(x) = f(f(\ldots f(x)\ldots)).$$

(ii) Let $g : R^1 \longrightarrow R^1$ be a differentiable increasing concave function. Then for $k \varepsilon R^1$

$$g^N(k) = \min_{y \varepsilon R^N} G(y;k) \tag{10}$$

and $\hat{y}_1 = g^{N-1}(k)$, $\hat{y}_2 = g^{N-2}(k)$, ..., $\hat{y}_{N-1} = g(k)$, $\hat{y}_N = k$ attains the minimum.

4. INVERSION, REVERSION AND CONJUGATION

First we consider the inverse function f^{-1} to a continuous strictly increasing function f. We remark that $f : R^1 \longrightarrow R^1$ is an onto differentiable strictly increasing convex function iff $f^{-1} : R^1 \longrightarrow R^1$ is the onto differentiable strictly increasing concave function. Then we have

COROLLARY (i) Let $f : R^1 \longrightarrow R^1$ be an onto differentiable strictly increasing convex function. Then for $k \varepsilon R^1$

$$f^{-N}(k) = \min_{y \varepsilon R^N} F^{-1}(y;k) \tag{11}$$

and $\hat{y}_1 = f^{-N+1}(k)$, $\hat{y}_2 = f^{-N+2}(k)$, ..., $\hat{y}_{N-1} = f^{-1}(k)$, $\hat{y}_N = k$ attains the minimum, where $F^{-1}(y;k)$ is the dynamic programmable function generated by f^{-1} and $f^{-n}(y)$ is the n-time composition of f^{-1}:

$$f^{-n}(y) = f^{-1}(f^{-1}(\ldots f^{-1}(y)\ldots)).$$

(ii) Let $g : R^1 \longrightarrow R^1$ be an onto differentiable strictly increasing concave function. Then for $h \in R^1$

$$g^{-N}(h) = \underset{x \in R^N}{\mathrm{Max}_N}\, G^{-1}(x;h) \tag{12}$$

and $x_1^* = g^{-N+1}(h)$, $x_2^* = g^{-N+2}(h)$, ..., $x_{N-1}^* = g^{-1}(h)$, $x_N^* = h$ attains the maximum, where $G^{-1}(x;h)$ is the dynamic programmable function generated by g^{-1}.

Here we remark that

$$F^{-1}(y;k) = F^{-1}(y_1) + f^{-1\cdot}(y_1)F^{-1}(y_2) + \ldots + f^{-1\cdot}(y_1)f^{-1\cdot}(y_2)$$
$$\times \ldots f^{-1\cdot}(y_{N-1})F^{-1}(y_N) + f^{-1\cdot}(y_1)f^{-1\cdot}(y_2)\ldots f^{-1\cdot}(y_N)k \tag{13}$$

where

$$F^{-1}(y) = f^{-1}(y) - yf^{-1\cdot}(y) \tag{14}$$

and $f^{-1\cdot}$ is the derivative of the inverse function f^{-1}. Similarly, $G^{-1}(x;h)$ is defined and omitted.

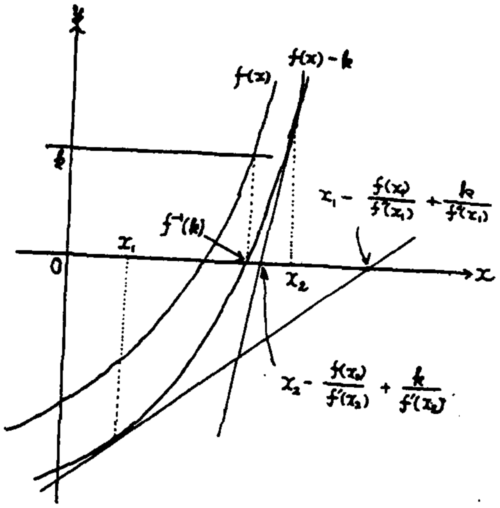
Second we consider the reversion of the linear approximation $f(x;h)$ of $f(x)$ —— not the reversion of $f(x)$ itself —— as follows. For any onto differentiable strictly increasing convex function $f : R^1 \longrightarrow R^1$, its linear approximation $f : R^1 \times R^1 \longrightarrow R^1$

defined by (2) or (3) is continuous strictly increasing and linear in h for $x \in R^1$. Therefore, $f(x;\cdot) : R^1 \longrightarrow R^1$ is invertible for $x \in R^1$. Its inverse function $f_{-1}(x;\cdot) : R^1 \longrightarrow R^1$ becomes

$$f_{-1}(x;k) = F_{-1}(x) + \frac{k}{f'(x)} \tag{15}$$

where

$$F_{-1}(x) = x - \frac{f(x)}{f'(x)}. \tag{16}$$



y
$f(x)$
$f(x)-k$
k
0
x_1
$f^{-1}(k)$
x_2
x
$x_1 - \frac{f(x_1)}{f'(x_1)} + \frac{k}{f'(x_1)}$
$x_2 - \frac{f(x_2)}{f'(x_2)} + \frac{k}{f'(x_2)}$

Fig. 2

We call $f_{-1} = f_{-1}(x;k)$ the *reverse function* of $f = f(x;h)$. As we noted in (1), we have

$$\begin{aligned} f(h) &= \max_{-\infty<x<\infty} f(x;h) \\ &= \max_{-\infty<x<\infty} [F(x) + f'(x)h] \qquad (17) \\ &= \max_{-\infty<x<\infty} [f(x) + (h-x)f'(x)] \end{aligned}$$

and $x^* = h$ attains the maximum. This fact is equivalently transformed to

$$\begin{aligned} f^{-1}(k) &= \min_{-\infty<x<\infty} f_{-1}(x;k) \\ &= \min_{-\infty<x<\infty} \left[F_{-1}(x) + \frac{k}{f'(x)}\right] \qquad (18) \end{aligned}$$

$$= \min_{-\infty<x<\infty} \left[x + \frac{k - f(x)}{f'(x)} \right]$$

and $x = f^{-1}(k)$ attains the minimum (see Fig.2). This fact reflects also the main idear of Newton method from a viewpoint of optimization. Therefore, we have the following reversed form of (9):

THEOREM 2. (i) Let $f : R^1 \longrightarrow R^1$ be an onto differentiable strictly increasing convex function. Then for $k \varepsilon R^1$

$$f^{-N}(k) = \min_{x \varepsilon R^N} F_{-1}(x;k) \tag{19}$$

and $\hat{x}_1 = f^{-N}(k)$, $\hat{x}_2 = f^{-N+1}(k)$, ..., $\hat{x}_{N-1} = f^{-2}(k)$, $\hat{x}_N = f^{-1}(k)$ attains the minimum, where $F_{-1} : R^N \times R^1 \longrightarrow R^1$ is the N-times composition of $f_{-1}(x;k)$:

$$F_{-1}(x;k) = f_{-1}(x_1;\ f_{-1}(x_2;\ \ldots;\ f_{-1}(x_N;k)\ldots)). \tag{20}$$

(ii) Let $g : R^1 \longrightarrow R^1$ be an onto differentiable strictly increasing concave function. Then for $h \varepsilon R^1$

$$g^{-N}(h) = \mathrm{Max}_{y \varepsilon R^N} G_{-1}(y;h) \tag{21}$$

and $y_1^* = g^{-N}(h)$, $y_2^* = g^{-N+1}(h)$, ..., $y_{N-1}^* = g^{-2}(h)$, $y_N^* = g^{-1}(h)$ attains the maximum, where $G_{-1} : R^N \times R^1 \longrightarrow R^1$ is the N-times composition of $g_{-1}(y;h)$:

$$G^{-1}(y;h) = g_{-1}(y_1;\ g_{-1}(y_2;\ \ldots;\ g_{-1}(x_N;h)\ldots)). \tag{22}$$

Here we remark that

$$F_{-1}(x;k) = F_{-1}(x_1) + \frac{F_{-1}(x_2)}{f'(x_1)} + \ldots + \frac{F_{-1}(x_N)}{f'(x_1)f'(x_2)\ldots f'(x_{N-1})}$$

$$+ \frac{k}{f'(x_1)f'(x_2)\ldots f'(x_N)} \qquad (23)$$

where $F_{-1}(x)$ is defined in (16). Similarly, $G_{-1}(y;h)$ is defined from $G_{-1}(y_n)$, $g'(y_n)$ and h. We call $F_{-1}(x;k)$, $G_{-1}(y;h)$ the *dynamic programmable function generated by reverse function $f_{-1}(x;k)$, $g_{-1}(y;h))$*, respectively.

We have the following relation between $F^{-1}(y;k)$ and F_{-1} $(x;h)$:

THEOREM 3. (i) Let $f : R^1 \longrightarrow R^1$ be an onto differentiable strictly increasing convex function. Then we have by the monotone transformation $y = f(x)$

$$f^{-1}(y;k) = f_{-1}(x;k). \qquad (24)$$

Furthemore, the monotone transfomation $y_n = f(x_n)$ $1 \le n \le N$ yields

$$F^{-1}(y;k) = F_{-1}(x;k). \qquad (25)$$

(ii) Let $g : R^1 \longrightarrow R^1$ be an onto differentiable strictly increasing concave function. Then we have by the monotone transformation $x = g(y)$

$$g^{-1}(x;h) = g_{-1}(y;h). \qquad (26)$$

Furthermore, the monotone transformation $x_n = g(y_n)$ $1 \le n \le N$ yields

$$G^{-1}(x;h) = G_{-1}(y;h). \tag{27}$$

Proof. It is straightforward.

Finally we consider conjugations $*$ and $\wedge$. For any convex function $f : R^1 \longrightarrow R^1$, we define its *conjugate function* $f^* : R^1 \longrightarrow R^1$

$$f^*(y) = \operatorname*{Sup}_{-\infty<x<\infty} [xy - f(x)]. \tag{28}$$

On the other hand, for any concave function $g : R^1 \longrightarrow R^1$, we denote its *conjugate function* $\hat{g} : R^1 \longrightarrow R^1$ by

$$\hat{g}(x) = \inf_{-\infty<y<\infty} [yx - g(y)]. \tag{29}$$

If both operations $*$ and $\wedge$ are well defined, they are dual in the following sense:

$$\widehat{(-f)}(y) = -f^*(-y) \qquad y \in R^1.$$

LEMMA 2. Let $f : R^1 \longrightarrow R^1$ be a twice differentiable strictly increasing strictly convex function. Then we have for $f'(-\infty) < y < f'(\infty)$

(i) $f^*(y) = xy - f(x)$

(ii) $f^{*\prime}(y) = x$ and in particular $f^{*\prime}(y) > 0$ for $f'(0)<y<f'(\infty)$

and

(iii) $f^{*\prime\prime}(y) = \frac{1}{f''(x)} > 0$

where x satisfies uniquely $f'(x) = y$. Therefore, $f^* : (f'(0),$

$f'(\infty)) \longrightarrow R^1$ is strictly increasing strictly convex. Thus we have the following result for f^*:

THEOREM 4. Let $f : R^1 \longrightarrow R^1$ be a twice differentiable strictly increasing strictly convex function. Then we have for $f'(0) < f^{*n}(h) < f'(\infty) \quad 0 \le n \le N-1$

$$f^{*N}(h) = \max_{\substack{f'(0)<y_n<f'(\infty) \\ 1\le n\le N}} F^*(y;h) \qquad (30)$$

and $y_1^* = f^{*(N-1)}(h)$, $y_2^* = f^{*(N-2)}(h)$, ..., $y_{N-1}^* = f^*(h)$, $y_N^* = h$ attains the maximum, where $F^*(y;h)$ is the dynamic programmable function generated by f^* and f^{*n} is the n-time composition of f^*.

Similarly, for concave function g, we have the following:

LEMMA 3. Let $g : R^1 \longrightarrow R^1$ be a twice differentiable strictly increasing strictly concave function. Then we have for $g'(\infty) < x < g'(-\infty)$

(i) $\hat{g}(x) = yx - g(y)$

(ii) $\hat{g}'(x) = y$ and in particular $\hat{g}'(x) > 0$ for $g'(\infty)<x<g'(0)$

and

(iii) $\hat{g}''(x) = \frac{1}{g''(y)} < 0$

where y satisfies uniquely $g'(y) = x$. Therefore, $\hat{g} : (g'(\infty),g'(0)) \longrightarrow R^1$ is strictly increasing strictly concave.

THEOREM 5. Let $g : R^1 \longrightarrow R^1$ be a twice differentiable strictly increasing strictly concave function. Then we have for $g'(\infty) < g^n(k) < g'(0) \quad 0 \le n \le N-1$

$$\hat{g}^N(k) = \min_{g'(\infty)<x_n<g'(0) \quad 1\le n\le N} \hat{G}(x;k) \tag{31}$$

and $\hat{x}_1 = g^{N-1}(k)$, $\hat{x}_2 = g^{N-2}(k)$, ..., $\hat{x}_{N-1} = g(k)$, $\hat{x}_N = k$ attains the minimum, where $\hat{G}(x;k)$ is the dynamic programmable function generated by $\hat{g}$ and $\hat{g}^n$ is the n-times composition of $\hat{g}$.

Here we remark that

$$F^*(y;h) = F^*(y_1) + f^{*\prime}(y_1)F^*(y_2) + \dots + f^{*\prime}(y_1)f^{*\prime}(y_2)\dots \times f^*(y_{N-1})F^*(y_N) + f^{*\prime}(y_1)f^{*\prime}(y_2)\dots f^{*\prime}(y_N)h \tag{32}$$

where

$$\begin{aligned} F^*(y) &= f^*(y) - yf^{*\prime}(y) \\ &= -f(x). \end{aligned} \tag{33}$$

Here x satisfies uniquely $f'(x) = y$. Similar expressions for $\hat{G}(x;k)$ and $\hat{G}(x)$ are omitted.

5. EXAMPLES

In this section we illustrate explicit form of $f(x;h)$, $F(x;h)$, $F^{-1}(y;k)$, $f_{-1}(x;k)$, $F_{-1}(x;k)$, $F^*(y;k)$ and others for a given $f(x)$.

5.1 $f(x) = e^x : (-\infty , \infty) \longrightarrow (0 , \infty)$

In this case we have the following expressions. First we have from (2),(6)

$$f(x;h) = (1 - x + h)e^{x} \qquad -\infty < x,h < \infty$$

$$F(x;h) = e^{x_1}(1 - x_1) + e^{x_1+x_2}(1 - x_2) + \ldots + e^{x_1+\ldots+x_{N-1}}$$

$$\times(1 - x_N) + e^{x_1+\ldots+x_N}\times h \qquad -\infty < x_n,h < \infty.$$

Second, for inversion, we have from (13),(14)

$$g(y) \equiv f^{-1}(y) = \log y : (0, \infty) \longrightarrow (-\infty, \infty) \qquad (34)$$

$$g(y;k) = f^{-1}(y;k) = -1 + \log y + \frac{k}{y} \qquad 0 < y,k < \infty$$

$$G(y;k) = F^{-1}(y;k) = -1 + \log y_1 + (y_1)^{-1}(-1 + \log y_2)$$

$$+ (y_1\ldots y_{N-1})^{-1}(-1 + \log y_N) + (y_1\ldots y_N)^{-1}k$$

$$y_n > 0, \quad k \gg 0$$

where $k \gg 0$ means that k is sufficiently large that log...logk (N-times log operation) becomes well defined. That is, in this case,

$$k > e^{e^{\cdot^{\cdot^{\cdot^{e}}}}} \qquad ((N-1)\text{'s } e).$$

Third, for reversion, we have from (15),(16),(20)

$$f_{-1}(x;k) = x - 1 + e^{-x}k \qquad -\infty < x < \infty, \quad k > 0$$

$$F_{-1}(x;k) = x_1 - 1 + e^{-x_1}(x_2 - 1) + \ldots + e^{-x_1-\ldots-x_{N-1}}(x_{N-1}-1)$$

$$+ e^{-x_1-\ldots-x_N}\times k$$

$$-\infty < x_n < \infty, \ k \gg 0.$$

Moreover, the reversion of $g = g(y)$ defined in (34) becomes

$$g_{-1}(x;h) = y(1 - \log y) + yh \qquad y > 0, \quad -\infty < h < \infty$$

$$G_{-1}(x;h) = y_1(1 - \log y_1) + y_1y_2(1 - \log y_2) + \ldots + y_1\ldots y_N$$

$$\times(1 - \log y_N) + y_1\ldots y_N h \qquad y_n > 0, \quad -\infty < h < \infty.$$

Fourth, for conjugation, we have from (28), (29), (32), (33)

$$f^*(y) = (-1 + \log y)y : (0, \infty) \longrightarrow [-1, \infty)$$

$$f^{*\prime}(y) = \log y > 0 \qquad \text{on} \quad (1, \infty)$$

$$f^{*\prime\prime}(y) = 1/y > 0$$

$$f^*(y;k) = -y + k\times\log y \qquad y > 1, \quad k > 1$$

$$F^*(y;k) = -y_1 - y_2\log y_1 - \ldots - y_N\log y_1\ldots\log y_{N-1}$$

$$+ k\times\log y_1\ldots\log y_N \qquad y_n > 1, \quad k > e^2$$

$$\hat{g}(x) = 1 + \log x : (0, \infty) \longrightarrow (-\infty, \infty)$$

$$\hat{g}(x;h) = \log x + x^{-1}h \qquad 0 < x,h < \infty$$

$$\hat{G}(x;h) = \log x_1 + (x_1)^{-1}\log x_2 + \ldots + (x_1\ldots x_{N-1})^{-1}\log x_N$$

$$+ (x_1\ldots x_N)^{-1}h \qquad x_n > 0, \quad h \gg 0$$

where $h \gg 0$ in this case means that

$$h > e^{-1+e^{-1+e^{\cdot^{\cdot^{-1+e^{-1}}}}}} \qquad (N\text{'s } e).$$

Finally, for reversion of $\hat{g}(y;k)$, we have

$$\hat{g}_{-1}(x;k) = -x\log x + xk \qquad x > 0, \quad -\infty < k < \infty$$

$$\hat{G}_{-1}(x;k) = -x_1\log x_1 - x_1x_2\log x_2 - \ldots - x_1\ldots x_N\log x_N$$
$$+ x_1\ldots x_N k \qquad x_n > 0, \quad -\infty < k < \infty.$$

5.2 $\quad f(x) = x^2 : [0, \infty) \longrightarrow [0, \infty)$

In this case we have the following result. First, we get

$$f(x;h) = -x^2 + 2xh \qquad x,h \geq 0$$

$$F(x;h) = -x_1^2 - 2x_1x_2^2 - \ldots - 2^{N-1}x_1\ldots x_{N-1}x_N^2$$
$$+ 2^N x_1\ldots x_N h \qquad x_n \geq 0, \ h \geq 0.$$

In particular Theorem 1 for case $N = 1$ implies

$$\underset{-\infty<x<\infty}{\text{Max}} \left[2xh - x^2 \right] = h^2 \qquad -\infty < h < \infty.$$

This is one of the simplest quasilinearization [1; p.134].

Second, the inversion becomes

$$g(y) \equiv f^{-1}(y) = \sqrt{y} : (0, \infty) \longrightarrow (0, \infty)$$

$$g(y;k) = f^{-1}(y;k) = \frac{1}{2}\left(y + \frac{k}{\sqrt{y}}\right) \qquad y,k > 0$$

$$G(y;k) = F^{-1}(y;k)$$
$$= \frac{1}{2}(y_1)^{1/2} + \frac{1}{2^2}(y_2/y_1)^{1/2} + \ldots + \frac{1}{2^N}(y_N/y_1\ldots y_{N-1})^{1/2}$$
$$+ \frac{k}{2^N}(y_1\ldots y_N)^{-1/2} \qquad y_n > 0, \ k > 0.$$

Therefore, Corollary (ii) for case $N = 1$ reduces

$$\min_{x>0}\left[\frac{1}{2}\sqrt{x} + \frac{1}{2\sqrt{x}}\right] = \sqrt{k} \qquad k > 0$$

(see also [1; p.134]).

Third, for reversion, we have

$$f_{-1}(x;k) = \frac{1}{2}(x + \frac{k}{x}) \qquad x,k > 0$$

$$F_{-1}(x;k) = \frac{1}{2}x_1 + \frac{1}{2^2}(x_2/x_1) + \dots + \frac{1}{2^N}(x_N/x_1\dots x_{N-1})$$

$$+ \frac{k}{2^N}(x_1\dots x_N)^{-1} \qquad x_n > 0, \quad k > 0.$$

Finally, the conjugation yields

$$f^*(y) = \frac{1}{4}y^2 : [0, \infty) \longrightarrow [0, \infty)$$

$$f^*(y;k) = -\frac{1}{4}y^2 + \frac{1}{2}yk \qquad y,k \geq 0$$

$$F^*(y;k) = -\frac{1}{4}y_1^2 - \frac{1}{4\cdot 2}y_1y_2^2 - \dots - \frac{1}{4\cdot 2^{N-1}}y_1\dots y_{N-1}y_N^2$$

$$+ \frac{1}{2^N}y_1\dots y_Nk \qquad y_n \geq 0, \quad k \geq 0$$

$$\overset{\frown}{g}(x) = -\frac{1}{4x} : (0\ ,\ \infty) \longrightarrow (-\infty\ ,\ 0)$$

$$\hat{g}(x) = -\frac{1}{2x} + \frac{h}{4x^2} \qquad x,h > 0.$$

Therefore we get

$$\hat{g}(x) = \min_{0<x<\infty} \overset{\frown}{g}(x;h)\ .$$

However if $N \geq 2$, then it does not hold that

$$\hat{g}^N(h) = \min_{0<x_n<\infty} \hat{G}(x;h) \qquad h > 0,$$

because of $\hat{g}(h) < 0$.

References

1. R. Bellman, Dynamic Programming, Princeton Univ. Press, Prinston N.J., 1957.
2. R. Bellman and R. Kalaba, Quasilinearization and Nonlinear Boundary-value Problems, American Elsevier, N.Y., 1965
3. R. Bellman and Wm. Karush, On a new functional transform in analysis : the maximum transform, Bull. Amer. Math. Soc. 67 (1961), 501-503.
4. R. Bellman and Wm. Karush, Mathematical programming and the maximum transform, J. SIAM Appl. Math. 10(1962), 550-567.
5. R. Bellman and Wm. KARUSH, On the maximum transform and semigroups of transformations, Bull. Amer. Math. Soc. 68(1962), 516-518.
6. R. Bellman and Wm. Karush, Functional equations in the theory of dynamic programming - XII: an application of the maximum transform, J. Math. Anal. Appl. 6(1963), 155-157
7. R. Bellman and Wm. Karush, On the maximum transform, J. Math. Anal. Appl. 6(1963), 67-74.
8. N. Furukawa and S. Iwamoto, Dynamic programming on recursive reward systems, Bull. Math. Statist. 17(1976) 103-126.
9. S. Iwamoto, Some operations on dynamic programmings with one-dimensional state space, J. Math. Anal. Appl. 69(1979), 263-282.
10. S. Iwamoto, Reverse function, reverse rpogram and reverse theorem in mathematical programming, J. Math. Anal. Appl. 95 (1983), 1-19.
11. S. Iwamoto, Sequential minimaximization under dynamic programming structure, J. Math. Anal. Appl. 108(1985), 267-282.
12. S. Iwamoto, R.J. Tomkins and C.-L. Wang, Some theorems on reverse inequalities, J. Math. Anal. Appl. 119(1986), 282-299.
13. E. Stanley Lee, Dynamic Programming, quasilinearization and dimensiomality difficulty, J. Math. Anal. Appl. 27(1968), 303-

322.

14. E. Stanley Lee, Quasilinearization and Invariant Imbedding, Academic Press, N.Y., 1968.

15. L.G. Mitten, Composition principle for synthesis of optimal multistage process, Operations Res. 12(1964), 601-619.

16. G.L. Nemhauser, Introduction to Dynamic Programming, John Wiely and Sons, 1966.

EXISTENCE AND COMPUTATION OF SOLUTIONS FOR THE TWO DIMENSIONAL MOMENT PROBLEM

György Sonnevend*

Inst. für Angewandte Mathematik,
Universität Würzburg
D-8700 Würzburg, Am Hubland

Introduction

In this paper we deal with some problems of the theory of two dimensional.polynomial moment problems. More precizely we give necessary and sufficient conditions for the existence of a solution, i.e. of a nonnegative mass distribution supported within a fixed, a priori given subset S of R^2, which has a finite set of moments with prescribed values.We study the problem of characterizing all minimal support solutions, i.e. those solutions which have a minimal number of atoms.

The connections between the restricted (or finite), classical, polynomial (onedimensional) moment problem (as a special case of the moment problems of Nevanlinna-Pick type) and various other problems in the theory of orthogonal polynomials, rational Pade approximation (interpolation of Stieltjes functions),restriction of self adjoint operators to Krylow-subspaces, construction of quadrature formulae, minimal partial realizations of causal linear input-output maps, are well known. Similar applications for the considered two dimensional generalization motivate our study. The method we use for the solution of these problems is operator theoretic and is based on solving an "extension problem" for pairs of commuting, self adjoint operators.The characterization obtained for the minimal support solutions,i.e. for the analogons of the Gaussian quadrature formulae is different from the previous approaches, which (as far as we know) used two dimensional orthogonal polynomials (searching for their common zeros) and polynomial ideal theory, see [11] for an extensive set of historical and current references. We were inspired by the operator theoretic treatment of moment problems as developped in [12], see also the method of the paper[16].

* on leave from Dept. of Numer.Anal.,Eötvös University
H. 1088,Budapest, Muzeum k.6-8,Föe'p.

Since the minimal support solutions are, in general non unique in the higher dimensional case (in contrast to the onedimensional case) moreover their set (thus the problem of f'nding at least one element of it) is not convex and for other reasons like the complexity and stability (with respect to errors in the prescribed moments)we propose and study here an other,particular (nonminimal) solution , i.e. mass distribution, the so called analytical centre of the feasible set (of solutions). Several positive features and applications of this solution concept,like stable computability with a relatively small number of arithmetical operations and the feasibility of high degree homotopy methods for computing bounds for any further,not specified "moment" (i.e. integrals with respect to the underlying measure)are studied in the last section.

2. Preliminaries

Suppose that $S \subset R^n$ is a closed set and μ is a nonnegative (Radon) measure supported within S. In the general, finite or restriced moment problem we shall study here the data are the N values reals

(2.1) $C_j = \int_S K_j(s)\ \mu(ds) = \varphi_j(\mu),\ j=1,\dots,N$

of fixed, linear (continuous) functionals φ_j, given by continuous on S functions K_j, j=1,...,N on S and one asks for the conditions of the existence and a characerization of all solutions μ which have minimal support belonging to S:

(2.2) $M \to \min,\ C_j = \Sigma K_j(s_k)\ \mu_k,\ \mu_k \geq 0,\ s_k \in S,\ k = 1,\dots,M.$

In the case when $S \subset R^2$, i.e. n=2, and for S = (x,y) the functions $K_1,\dots,K_N$ have the form

(2.3) $x^i y^j,\ (i,j) \in I,\ |I| = N$

where I is a finite subset of Z_+^2 (the set of nonnegative entires)of cardinality N, the above problem - the so called restricted polynomial moment problem - is a natural generalization of the Gaussian quadrature problem. Of course, one can expect a reasonably simple and constructive answer to this problem only if I and S have a simple form, e.g. S is a quadrangle

(2.4) $S = [a_1,b_1] \times [a_2,b_2\ ]$

and - for some fixed, positive L -

(2.5) $I = \{(i,j) \mid i + j \le L,\ i,j \ge 0\}$.

We give now an equivalent formulation of the problem(2.2)-(2.3) which is crucial for our approach.

Proposition 1. The problem (2.2)-(2.3)- with data set[c(I),S]is equivalent to the existence and characterization of quadruples H,A,B,e , where H is a Hilbert space (whose dimension should be minimized), A and B are self adjoint cummuting operators on H and e is a nonzero vector in H such that

(2.5) $c_{ij} = \langle A^i B^j e,e \rangle$, for all $(i,j) \in I$.

Proof. If there is a solution of problem (2.2)-(2.3) then we define the Hilbert space

(2.6) $H: = L_2(S,d\mu)$, $e: = 1$ on S

and the operators

(2.7) $A\, f(x,y) := x\, f(x,y)$ $\quad B\, f(x,y): = y\, f(x,y)$

which are self adjoint and commuting. The conditions in (2.2)can be expressed as those in (2.5).

Conversely, suppose that (2.5) holds and let A,B have the eigenvectors (they are common and form a basis of H by the communtativity and self adjointness of A,B) $\Psi_1,\ldots,\Psi_M$and eigenvalues $x_1,\ldots,x_M$ resp. $y_1,\ldots,y_M$,where M is the dimension of H

(2.8) $A\Psi_k = x_k \Psi_k$, $B\Psi_k = y_k \Psi_k$, $k= 1,\ldots,M$.

Then

(2.9) $c_{ij} = \sum_{k=1}^{M} x_k^i y_k^j \mu_k$, $(i,j) \in I$, where $\mu_k: = \langle \Psi_k,e\rangle^2$, $k=1,\ldots,M$.

This completes the proof and shows that once we constructed the quadruple <H,A,B,e> then the quadrature formula (2.9)can be obtained by a low complexity stable numerical method i.e.solving an eigenvalue problem.

Not assuming H to be finite dimensional we had to invoke the general spectral decomposition theorem, see e.g. [12] ,by which a representing measure is obtained from the associated projector measure

$$d\mu(\lambda) = d(\langle E(\lambda)e,e\rangle)$$

Proposition 2. If problem (2.1),(2.3) has a solution then the problem (2.2),(2.3) also has a solution,moreover for the minimal value M we have the inequality

(2.10) $\min M \le |I|$

which is exact in the sense, that there exist (multiple connected) domains S such that for the constant weight function $\mu'(x,y) \equiv 1$ on S and the set I as in (2.5), for arbitrary L we have equality in (2.10) - The first part is known as Chakaloff's theorem see [6] and is based on the simple fact that if

$$C = \sum_{i=1}^{R} \gamma_i e_i, c \in R^k, \ \gamma_i \geq 0, i=1,\ldots,R$$

then there exist a similar representation in which there are at most k nonzero constants γ_i. For a proof of the second part see §4,ch.2in [11]. Before going further let us indicate here the connection of the above problem with the minimal,partial relization problem for a class ot two dimensional shift invariant, linear input-output maps

(2.11) $y_{k,l} = \sum_{k\geq i, l\geq j} F_{k-i,l-j} U_{i,j}$

by state-space models of the form

(2.12) $y_{k,l} = <h, x_{k,l}>$

$$x_{k+1,l+1} = F_1 x_{k,l+1} + F_2 x_{k+1,l} - F_1 F_2 x_{k,l} + gU_{k,l}$$

where F_1, F_2 are commuting,symetric matrices in R^M and $h,g \in R^M$,see [3]. The transfer functions assiciated to such maps

$$T(w,z) = \iint \frac{d\mu(x,y)}{(1-wx)(1-zy)} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} F_{ij} w^i z^j$$

are generalizations of the one variable Stieltjes functions and should play the same role in analyzing "passive" input-output maps. Note that the realizability conditions have the form of complete, infinite moment conditions, if $g = h$,

$$F_{i,j} = <h, F_1^i F_2^j g> , \quad i,j \geq 0.$$

It is known that the minimal partial realization problem underlies most of the basic engineering problems of system analysis, see e.g. [2], even if for a suitable,more exact and stable numerical solution of these problems other linear information functionals are better suited, see [14] and below.Connections to(rational)approximation (interpolation) problems for Stieltjes functions are extensively studied, see e.g.[7], [10],[14],[16].

3.Exact conditions of existence and minimality

We shall restrict our interest to so called "regular"index sets I, which - by definition- have the following property.

(3.1) if $(i,j) \in I$, then $(k,l) \in I$, for all $k \leq i$, $l \leq j$

In order to characterize the minimal solutions (H,A,B,e) we have to characterize first the sets with consits of a maximal number of linearly independent vectors among(3.2) $A^i B^j e$, $i,j \geq 0$.

Lemma 1 In the linear space H spanned by the vectors (3.2) (if it is finite dimensional) there always exist a basis consisting of elements of a regular subset $L \subset Z_+^2$.

Proof. Let n_1 be the maximum of the values n such that $e, Be, .., B^n e^{-1}$ are linearly independent. Suppose inductively that $n_k, k \geq 1$ is the largest value of n such that $B^{n-1}A^{k-1}b$ is linearly independent on tne vectors A^iB^je, with $i \leq k-2, j \leq n_i$ and $i=k-1, j \leq n-2$. Since the sequence of the $n_k, k-1, \ldots$ satisfies $n_1 \geq n_2 \ldots \geq n_k \geq 1$, $\sum_k n_k = \dim H$ the above procedure ends in at most dim H steps and yields a regular set L.

Definition. If L is a regular set, the (generalized) Hankel-matrix associated to it is defined by

$$H_{L(i_1,j_1),(i_2,j_2)} := C_{i_1+i_2, j_1+j_2}$$

where we order the rows and colums of H_L (indexed by elements of L) according to the lexigographic order in Z_+^2. Further we denote-for a regular set L

$$L^*: = \{(k,l) \mid \exists\ (i,j) \in L \text{ with } 1 \geq k-i \geq 0,\ 1 \geq l-j \geq 0\}$$

$$L_1 := \{(k,l) \mid \exists (i,j) \in L, k \leq i+1, l=j\}, L_2 = \{(k,l) \mid \exists (i,j) \in L, k=i, l \leq j+1\}$$

$$L^2 := \{(k,l) \mid k=i_1+i_2, l=j_1+j_2, (i_1,j_1) \in L, (i_2,j_2) \in L\}$$

Theorem 1. The necessary and sufficient condition for the existence - given the moment data c(I) - of a nonnegative representing measure supported in at most M points of S is that there exists a regular set L of cardinality m and an extension of the data from c(I) to $c((L^*)^2)$, i.e. an assigument of values to the unspecified moments in $c((L^*)^2)$ such that the matrix $H_L{}^*$ is positive semidefinite and

(3.3) $\operatorname{rank} H_L = \operatorname{rank} H_L{}^* \leq M$.

Moreover the minimal value of M for which the above two conditions can be satisfied equals the minimal number of knots in the corresponding cubature formula.

Proof. In order to understand the role of the matrices H_L and $H_L{}^*$ note that these are the Gram matrices associated to the set of vectors

$$W(L) = \{A^iB^je \mid (i,j) \in L\}$$

$$W(L^*) = \{A^rB^sv \mid v \in W(L),\ 0 \leq r \leq 1,\ 0 \leq s \leq 1\}.$$

The necessity of the conditions (3.3) follows now from Proposition 1 and Lemma 1 since Gram matrixes should be positive semidefinite and their rank equal the dimension of the space spanned by the underlying vectors. To prove the sufficiency of the conditions we have to construct a quadruple (H,A,B,e) , such that dim H = rank H_L and (2.5) holds. Now we define H as the Hilbert space spanned by vectors V_{ij} indexed by the element $(i,j) \in L^*$, whose scalar products are specified by

$$\langle V_{i,j}, V_{k,l} \rangle = c_{i+k,j+l}$$

Since rank H_L = rank H_L^* , the operators A,B defined by

$$AV_{i,j} = V_{i+1,j} \quad BV_{i,j} = V_{i,j+1}, \quad (i,j) \in L$$

are hereby defined on the whole space H, moreover they are well defined: if

$$V_{r,s} = \sum_{(i,j)\in L} \alpha_{i,j} V_{i,j}, \text{ i.e. } \sum \alpha_{ij} \langle V_{i,j}, V_{k,l} \rangle = \langle V_{r,s}, V_{k,l} \rangle$$

for all $(k,l) \in L^*$, then

$$V_{r+1,s} = \sum_{(i,j)\in L} \alpha_{i,j} V_{i+1,j} \text{ and } V_{r,s+1} = \sum_{(i,j)\in L} \alpha_{i,j} V_{i,j+1}$$

hold. Indeed multiplying the latter relations by $V_{k,l}$, $(k,l) \in L$, the relations obtained are consequences of the previous ones because H_L is a submatrix of H_{L_1} and H_{L_2} and these are submatrixes of H_L*.

These operators A and B are clearly symmetric (i.e. self adjoint) since for all $(i,j),(k,l) \in L$

$$\langle AV_{i,j}, V_{k,l} \rangle = c_{i+1+k,j+l} = \langle V_{i,j}, AV_{k,l} \rangle$$

and they commute, since

$$\langle ABV_{i,j}, V_{k,l} \rangle = c_{i+k+1,j+l+1} = \langle BAV_{i,j}, V_{k,l} \rangle .$$

By this the theorem is proved.

The difficulty with this extension problem is partly apparent from the following fact:the restriction of the original say infinite dimensional operators A and B to a Krylow-like subspace W(L) are symmetric but they may not commute, (in general, they do not commute). It is not clear what further connections (if any) exist between the set I (and the values c(I)) on one side and the possible sets L on the other side, is it true that L can be chosen as a subset of I?

These sharp differences between one and higher dimensional polynomial moment problems have been observed e.g. in [13],where it is first shown that in the twodimensional trigonometric,finite moment problem the nonnegativity of the associated, generalized Toeplitz matrix (the precize analogon of our Hankel matrix) is not sufficicient for the solvability. The theory of normal extensions of operators, see the appendix written by Szőkefalvi Nagy in [12], is clearly related to our problem since the operator A + iB = T should be normal, for A,B to be symmetric and commuting and vice versa. The conditions - in terms of c(I) - for the condition: spectrum $T \subset S$ can be easily written down in the case (2.4): the following matrixes should be nonnegative definite

(3.4) $H_{L_1} - a_1 H_1$, $b_1 H_L - H_{L_1}$, $H_{L_2} - a_2 H_L$, $b_2 H_L - H_{L_2}$.

If S is the disjoint union of two quadrangles Q_1 and Q_2 than we have to require that there exist a decomposition of each of the moments (fixed or assigned) such that

$$c_{i,j} = c^{1}_{ij}\ (Q_1) + c^{2}_{i,j}(Q_2)\ ,\ (i,j) \in (L^*)^2$$

and (3.4) holds for the respectively decomposed matrixes.As an example of a simple application of Theorem 1 we metion the following fact: fir six data $(c_{o,o};c_{1,o};\ldots\ ;c_{o,2})$ if the coresponding 3 x 3 matrix is nonsingular the minimal measures should have 3 atoms and they constitute a one parameter family.

A new numerical approach to solve the existence problem

It is very difficult to handle the constraint(3.3) numerically, the set of solutions of the minimal extension problem is not convex. Observing that the finite dimensional analgon of the solution set to a moment problem (2.1) has the form of a polyhedron (in the sequel we often use abbreviations for N tuples $(c_1,\ldots,c_N) = c^N$)

(4.1) $K = K\ (k^N,c^N) = \{\ \mu\ |< k_i,\mu>= c_i,\ i=1,\ldots,N,\ \mu \in R^m_+\ \}$

we see that searching for the extremal points "vertices" of K. It is known that the parameters of a Gaussian quadrature are very ill conditioned functions of the moments (note that(2.1) is something like an integral equation of the first order whose right hand side is known only at some points) - and this has its parallel in the fact that the vertices of a polyhedron $H(k^N,\ c^N)$ are nonsmooth functions of the data c^N, or (k^N,c^N).

We propose now using an other, specific solution, the "analytic centre" of the solution set, in order to solve the existence (and some related estimation) problems, in a numericaly more feasible manner.

The analytic centre $\mu(K) = \mu(k^N,c^N)$ of the polyhedron(4.1) is defined as the unique point which solves the following optimization problem

$$\max\{\sum_{i=1}^{m}\log\ \mu_i|\langle k_j,\mu\rangle= c_j, j=1,\ldots,N, \mu_i\ \geq 0, i=1,\ldots,m\}$$

If the polyhedron is represented in its own space(of dimension m-N,in general), i.e. $K \to P = P(a^m,b^m)$

$$P(a^m,b^m) = \{x\,|\,b_i-\langle a_i,x\rangle \geq 0,\ i=1,\ldots,m, x \in R^{m-N}\}$$

by the map $\mu_i = b_i-\langle a_i,x\rangle$,$i=1,\ldots,m$, then $\bar{\mu} = \bar{x}\ (a^m,b^m)$ the point, which solves the problem (assuming int $P \neq \emptyset$)

$$\max\{\prod_{i=1}^{m}\ (b_i-\langle a_i,x\rangle)\,|\,x \in P(a^m,b^m)\}.$$

One can prove that the map $(a^m,b^m) \to \bar{x}(a^m,b^m)$ is affine invariant and there exists a two sided ellipsoidal approximation for P around $\bar{x}$:

$$\bar{x} + E \subset P \subset \bar{x} + m\ E,\ E =\{\ z\,|\langle Az,z\rangle \leq 1\}$$

where the symetric matrix $E = E(a^m, b^m)$ is easily obtained from $\bar{x}(a^m, b^m)$ see [14],[15]. The fact that $\bar{x}(a^m, b^m) = \bar{\mu}(k^N, c^N)$ is an analytic, very smooth function of the data allows to solve the feasibility and linear optimization problems by a homotopy approach, see [15], which we generalize now as follows.

The analytic entre of the set (2.1) is defined (if its exists) as the solution of the problem

(4.2) $\sup\{\int_S \log \mu'(s)ds \mid \int_S K_j(s)\mu'(s)ds = c_j,\ j=1,\dots,N\}.$

It is easy to prove that the set of values c^N for which (4.2) has a solution is convex and dense in the set of all feasible c^N, if S is a domain, i.e. closure (int S) = S. For the trigonometric moment problem this solution was studied already about 1920, see [10],[14].

Lemma 1. The solution of the problem (4.2) - if it exists - has the following form

$$\mu'(s) = \left(\sum_{j=1}^{N} \alpha_j K_j(s)\right)^{-1}$$

for suitable $\alpha^N \in R^N$, which in fact is then the unique solution of the equation

(4.3) $$\frac{\partial F(\alpha)}{\partial \alpha_j} = \int_S K_j(s) \left(\sum_{j=1}^{N} \alpha_j K_j(s)\right)^{-1} ds = c_j, j = 1,\dots,N$$

such that $\Sigma\, \alpha_j K_j(s)$ is positive on S, here

(4.4) $$F(\alpha) = \int_S \log \left(\Sigma\, \alpha_j K_j(s)\right) ds$$

Proposition The moment problem (2.1) has a solution if and only if the homotopy path $\alpha^N(\lambda)$ can be continued from $\lambda = 1$ till $\lambda = 0$, where $\alpha^N(\lambda)$ $0 < \lambda \leq 1$ is defined as the solution of (4.3) where c^N is replaced by $(1 - \lambda)c^N + \lambda c_o^N$,

$$c_o^N = \int_S K^N(s) \left(\sum_{j=1}^{N} \overset{o}{\alpha}_j K_j(s)\right)^{-1} ds$$

and $\sum_{j=1}^{N} \overset{o}{\alpha}_j K_j$ is an arbitrarily fixed polynom which is positive on S. The proof is a simple application of the implicite function theorem. For brevity we can only refer to [9], [15] for the application of this method for the estimation of (computation of exact upper and lower bounds in terms of the moments c^N for

$$l(c^N) \leq \int K_o(s)\, \mu(ds) \leq u(c^N)$$

It can be expected that for smooth analytic kernel functions $K_o, K_1, \dots, K_N$ this approach is superior to those using discretizations of the measure (of the set S) and algorithms based on the simplex method (note that the latter methods use- as a tool - extremal solutions, only piecewise smooth homotopies); concerning numerical test results on this approach-using homotopies along analytic centers- to solve linear programming problems, see[9].

The special solution of (4.2) in the case of the (real)trigono - metric moment problem - where $K_j(s) = \exp(i(j-1)s)$, $s \in [-\Pi,\Pi]$ and μ a measure on $[-\Pi,\Pi]$ (which is symetrical to zero)-,which is a special case of the Nevanlinna-Pick moment problem, is the so called "maximum entropy" solution. These analytical centers, more precizely the coefficients of the trigonometric polynomial $[\bar{\mu}'(e^{is})]^{-1}$ are rational functions of c^N which can be computed rather quickly:in $O(N^2)$ arithmetical operations. This and other observations, see[15],lead to the idea that for the extrapolation of the function $\alpha^N(\lambda)$rational (multipoint Pade) approximation - with Newton type corrector step to solve (4.3) - will furnish a rather efficient path following method. In fact, in a problem closely related to (4.2), the use of a special rational extrapolation method can be justified rigorously using a generalizaiton of the well knwon fact (see e.g.[7]) that the multipoint Pade approximants (i.e.interpolants) to a Stieltjes function are again Stieltjes functions, see[15].

In order to solve - over some domain S - the closely related uniform approximation problems

$$\min_{\beta^N} \| K_o(s) - \sum_{i=1}^{N} \beta_i K_i \|_{L^\infty(S)}$$

we propose following the homotopy path $\beta^N(\dot{\lambda})$ determined by

$$\sup_{(\varepsilon,\beta^N)} (\log(\lambda-\varepsilon) + \int_S (\log (K_o(s)-\sum_{i=1}^{N} \beta_i K_i(s)-\varepsilon)+ \log (\varepsilon-K_o(s)+ \sum_{i=1}^{N} \beta_i K_i(s)))ds .$$

Of course,the sucess of these methods depends (among others) on the availability of fast and accurate methods for approximating the above integrals as well as those in (4.3).

References

[1] T.Ando,Truncated moment problems for operators,Acta Sci.Math. (Szeged),31 (1970), 319-334.

[2] A.Antoulas, A New Approach to Synthesis Problems,IEEE Trans.Ant. Contr.,vol.30,No.5(1985)465-473.

[3] S.Attasi, Modelling and Recursive Estimation for Double Indexed Sequences, in System Indentification,ed. by R.K.Mehra and D.G.Lainiotis,Academic Press, 1976,pp.289-348.

[4] M.F.Barnsley and P.D.Robinson, Rational Approximant bounds for a class of two variable Stieltjes functions,SIAM J.Math.Anal. vol.9,No 2, 1978,pp.272-290.

[5] N.K.Bose,(editor),Multidimensional System Theory,D.Reidel, Dordrecht,1985.

[6] V.Chakaloff, Formules de cubatures mecaniques a coefficients non nêgatifs, Bull.Sci.Math.,Ser.2. 1957,81.No 3,123 - 134.

[7] G.Cybenko, Restrictions of normal operators, Pade approximation and antoregressive time series, SIAM J.Math.Anal. 15 (1984), 753 - 767.

[8] A.A. Goncar,G.Lopez, On Markov's theorem for multipoint Pade approximation,Math. USSR-Sb.,vol 34 (1978),449-459.

[9] F.Jarre,G.Sonnevend,J.Stoer, An Implementation of the method of analytic centers, Report No. 34, Schwerpunktprogramm der DFG, Anwendungsbezogene Optimierung und Steuerung, Univ.Würzburg, 16 p.,appears in Proc. 8th INRIA Conf. on Analysis and Optimization of systems, Antibes,1988.

[10] M.G.Krein, A.A.Nudelman,Markov's Moment Problem and Extremal Problems (in russian) Nauka,Moscow,1973.

[11] I.P.Mysovskih, Cubature formulae (in russian),Nauka,Moscow,1981.

[12] F.Riesz,B.Szökefalvi Nagy,Lecons D'Analyse Fonctionelle,Akad.Budapest,1972.

[13] W.Rudin,Function theory in polydiscs, W.A.Benjamin Inc. New York ,1969.

[14] G.Sonnevend, Sequential, stable and low complexity methods for the solution of moment (mass recovery)problems, in Proc.4. Int:Conf. on Numerical Methods Colloquia Societatis J.Bolyai, vol.50,pp.635-668, North Holland,Akademia,Budapest,1987.

[15] G.Sonnevend, J.Stoer,Global Ellipsoidal Approximations and Homotopy Methods for solving convex, analytic programms, Schwerpunktprogramm der DFG, Report No. 4, Jan. 1988, Submitted to Numerische Mathematik.

[16] Szökefalvi Nagy,B.,A.Korányi,Operatorentheoretische Behandlung und Verallgemeinerung eines Problemskreises in der komplexen Funtkionetheorie, Acta Mathematica,vol. 100(1958),pp.171-202.

AN APPROXIMATION PROCEDURE FOR STOCHASTIC CONTROL PROBLEMS. THEORY AND APPLICATIONS

Roberto Gonzâlez and Edmundo Rofman

Facultad de Ciencias Exactas e Ingeniería, Av. Pellegrini 250, (2000) Rosario, Argentina.
INRIA, Domaine de Voluceau, BP 105, Rocquencourt, 78153 Le Chesnay Cedex, France.

ABSTRACT

The aim of this paper is to propose an approximation procedure to compute the value function V and the optimal policy $\hat{u}$ related to the stochastic problem $(\mathcal{P})$ of controlling diffusion processes. This procedure can be easily extended to problems for which stopping time and impulse controls are also considered.

O - INTRODUCTION

As we did in [8] for deterministic problems we will employ here as basic tool of analysis the characterization of V as the maximum element of a suitable set $\mathcal{W}$ of functions w. While in [8] the definition of $\mathcal{W}$ requires for w to be subsolution of the first order Hamilton-Jacobi-Bellman equation, i.e. :

$$\frac{\partial w(x)}{\partial x} \cdot f(x,u) + \ell(x,u) - \alpha w(x) \geqslant 0, \forall u \in U, \tag{0.1}$$

here, in the stochastic case, we deal instead of (1) with

$$\mathcal{L}(u)w + \ell(u) \geqslant 0 \tag{0.2}$$

where $\mathcal{L}$ is a second order differential operator.

In what follows $(\mathcal{P})$ will be solved using the characterization mentioned above. To introduce the discretized problems $(\mathcal{P}^h)$ we need to define properly the functions w^h belonging to $\mathcal{W}^h$. In fact : the existence of maximum solution V^h for each problem $(\mathcal{P}^h)$ and the convergence of V^h to V are shown using a Discrete Maximum Principle (DMP) that w^h must verify (cfr. [3]). To insure this property we use particular schemes to discretize the first and second derivatives of w. Furthermore this choice enable us to compute V^h using an algorithm of relaxation type that increases the values of w^h in the vertices of the triangulation employed.

Comments on applications are included in the final chapter.

1 - THE PROBLEM (P)

Let us consider :

a) The complete probabilistic space

$$(\Omega, P, \mathcal{F}, \mathcal{F}(t)) \ ; \tag{1.1}$$

b) The state process y(.), modelled by the diffusion

$$\begin{aligned} dy(t) &= f(y(t), u(t))dt + \sigma(y(t), u(t))\, dw(t) \\ y(o) &= x \ , t \geqslant 0, y \in Q \subset R^n \end{aligned} \tag{1.2}$$

with

Q : open boundet set
w(t) : Wiener process $\mathcal{F}(t)$-measurable
u(t) : control process progressively measurable in a compact set $U \subset R^m$
σ is a n $\times$ n matrix
f and σ bounded continuous on $Q \times U$.

c) The cost functional

$$J(x,u(.)) = E \left\{ \int_0^\tau \ell(y(s).u(s))\, e^{-\alpha s}\, ds \right\} \tag{1.3}$$

with

τ : first exit time of $\overline{Q}$ of the system trajectory
$\alpha > 0$
ℓ : bounded continuous function on $Q \times U$.

Let us introduce the definition of the optimal cost

$$V(x) = \inf_{u \in U} J(x,u(.)), \tag{1.4}$$

V(x) being solution (cfr. [5],[2]) of the Hamilton-Jacobi-Bellman equation

$$\begin{aligned} &\min_{u \in U} \{\mathcal{L}(u)V + \ell(.,u)\} = 0 \text{ in } Q \\ &V = 0 \text{ in } \partial Q \end{aligned} \tag{1.5}$$

where the differential operator $\mathcal{L}$ is given by :

$$\mathcal{L}(u) = \sum_{r,s=1}^{n} a_{rs}(x,u) \frac{\partial 2}{\partial x_r \partial x_s} + \sum_{r=1}^{n} f_r(x,u) \frac{\partial}{\partial x_r} - \alpha \tag{1.6}$$

with

$$a_{rs} = \tfrac{1}{2} \sum_{z=1}^{n} \sigma_{rz}\, \sigma_{zs}, \text{ i.e. } a_{rs} = a_{sr}. \tag{1.7}$$

As it was said in the Introduction we will compute V taking advantage of its characterization as maximum element of a suitable set, i.e. (cfr. [6], [8], [15]) solving the following auxiliar problem (having V as solution) :

$(\mathcal{P})$: Find the maximum element $\tilde{w}$ of the set

$$\mathcal{W} = \{w \in W_0^{1,\infty}(Q) \; / \; \mathcal{L}(u)w + \ell \geq 0 \text{ in } \mathcal{D}'(Q) \; \forall u \in U, \; Q \subset \mathbb{R}^n\} \tag{1.8}$$

being

$$w \leq \tilde{w} \Leftrightarrow w(x) \leq \tilde{w}(x), \; \forall x \in Q \tag{1.9}$$

the natural partial order in $\mathcal{W}$.

(Questions concerning existence and unicity of the solution of $(\mathcal{P})$ can be seen in [4], [15]).

2 - THE DISCRETIZED PROBLEM ($\mathcal{P}^h$)

2.1. Preliminary comments

We will compute V as the limit of the solutions of a sequence of approximate problems ($\mathcal{P}^h$).

To simplify the presentation we will suppose that Q is polyhedric. We consider in Q a triangulation Q^h (union of simplices), x_i^h being ($i = 1, 2, ..., N_h$) the vertices of Q^h.

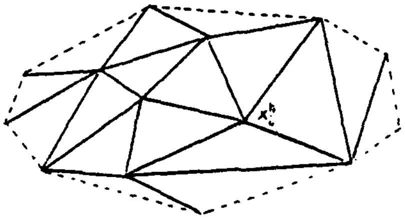


Fig 1

Then we define $\mathcal{W}^h$ by functions w^h verifying properties related to (1.8), (1.6). The main difficulty of this approach is to ensure the existence of a maximum element $\overline{w}^h$ in $\mathcal{W}^h$.

Following what we did in [8] for the deterministic case we introduce in $\mathcal{W}^h$ the natural partial order

$$w_1^h \leq w_2^h \Leftrightarrow w_1^h(x_i^h) \leq w_2^h(x_i^h),\ \forall x_i^h \text{ vertex of } Q^h \tag{2.1}$$

We consider functions $w^h : \overline{Q}^h \to \mathbb{R}$, w^h continuous in $\overline{Q}^h$ with $\frac{\partial w^h}{\partial x}$ constant in the interior of each simplex of Q^h, i.e., w^h are linear finite elements. So, to define w^h it will be enough to precise the inequality ("discretization" of $\mathcal{L}(u)w + \ell \geq 0$) to be verifyied at each vertex x_i^h of Q^h. Taking [8] into account if suffices to propose a suitable discretization of $L(u)w = \sum_{r,s=1}^{n} a_{rs} \frac{\partial^2 w}{\partial x_r\, \partial x_s}$, the term containing the second order derivatives of w.

2.2. Definition of $L^h(u)\, w^h$

Let us consider $S(x_i^h)$ (see Fig. 2), all the simplices having x_i^h as vertex.

From (1.7) the matrix $A = (a_{rs})$ has no negative eigenvalues λ_p and orthogonal eigenvectors. So

$$A = UDU' \tag{2.2}$$

with $UU' = I$

D (diagonal) / $D_{pp} = \lambda_p \geq 0.$

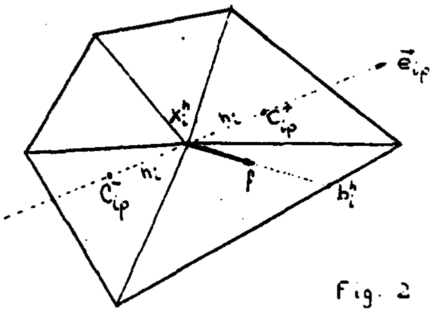


Fig. 2

If we consider, with center in x_i^h a new coordinates system (we denotes G the transformation matrix $G(x_i^h)$:

$$\eta = G \, . \, \xi \tag{2.3}$$

and we define

$$w(x) = w(x_i^h + \xi) \overset{\Delta}{=} \tilde{w}(\xi) = \tilde{w}(G^{-1}\eta) \overset{\Delta}{=} \hat{w}(\eta) \tag{2.4}$$

we obtain

$$L(u)w = \sum_{r,s=1}^{n} a_{rs}(x_i^h, u) \frac{\partial^2 w}{\partial x_r \, \partial x_s} = \sum_{p,q=1}^{n} b_{pq}(x_i^h, u) \frac{\partial^2 \hat{w}}{\partial \eta_p \, \partial \eta_q}, \tag{2.5}$$

with $b_{pq}(x_i^h, u) = (GAG')_{pq}$.

So, after the choice $G = U'$ we have, because

$$b_{pq} = \lambda_p \, \delta_{pq} \tag{2.6}$$

the following diagonal form of L :

$$Lw = \sum_{p=1}^{n} \lambda_p(x_i^h, u) \frac{\partial^2 \hat{w}}{\partial \eta_p^2} \tag{2.7}$$

Now we define naturally the approximated oeprator L^h :

$$L^h\, w^h(x_i^h) = \sum_{p=1}^{n} \lambda_p(x_i^h,u)\left(\frac{\partial 2}{\partial \eta_p^2}\right)^h \hat{w}^h \qquad (2.8)$$

where $\left(\frac{\partial 2}{\partial \eta_p^2}\right)^h \hat{w}^h = \frac{1}{h_i^2}\,(\hat{w}^h(C_{ip}^-)_\eta - 2\,\hat{w}^h(x_i^h)_\eta + \hat{w}(C_{ip}^+)_\eta)$ with

$$C_{ip}^- = x_i^h - h_i\,\vec{e}_{ip} \qquad (C_{ip}^-)_\eta = (0,0, ..., -h_i, ..., 0)$$

$$C_{ip}^+ = x_i^h + h_i\,\vec{e}_{ip} \qquad (C_{ip}^+)_\eta = (0,0, ..., h_i, ..., 0)$$

$$(x_i^h)_\eta = (0,0, ..., 0, ..., 0)$$

$\vec{e}_{ip}$ giving the direction of the η_p-axis and h_i such that $C_{ip}^-, C_{ip}^+ \in S(x_i^h)$.

2.3. Definition of $\mathcal{W}^h$

Coming back to (1.6), $\sum_{r=1}^{n} f_r(x,u)\,\frac{\partial}{\partial x_r}$ will be discretized as it was done in [8], i.e., we will consider *∇ in the direction f* (see Fig. 2) :

$$f\,.\,\nabla\, w^h(x_i^h) = \frac{w^h(b_i^h) - w^h(x_i^h)}{\| b_i^h - x_i^h \|}\,.\,\| f(x_i^h) \|\,. \qquad (2.9)$$

So, from (2.8) and (2.9) we can define

$$\mathcal{W}^h = \{w^h : Q^h \to R \;/\; L^h(u)w^h + \ell(u) \geqslant 0,$$

$$\forall u \in U^h,\ \forall x_i^h \in Q^h,\ w^h \leqslant 0 \text{ on } \partial Q^h\} \qquad (2.10)$$

where U^h is a finite discretization of U and

$$\mathcal{L}^h(u)w^h(x_i^h) + \ell(u, x_i^h) = \sum_{p=1}^{n} \frac{1}{h_i^2} \lambda_p(x_i^h,u)(w^h(x_i^h - h_i \vec{e}_{ip}) - 2\, w^h(x_i^h)$$

$$+ w^h(x_i^h + h_i \vec{e}^{\,+}_{ip})) + \frac{w^h(b_i^h) - w^h(x_i^h)}{\| b_i^h - x_i^h \|} \| f(x_i^h) \| - \alpha\, w(x_i^h) + \ell(u, x_i^h). \qquad (2.1,1)$$

Finally we can consider the discretized problem $(\mathcal{P})^h$: Find the maximum element $\overline{w}^h$ of the set $\mathcal{W}^h$ with respect to the partial order (2.1), i.e. find $\overline{w}^h(x)$ such that $\overline{w}^h(x_i^h) \geq w(x_i^h)$, $\forall x_i^h \in Q^h, \forall w^h \in \mathcal{W}^h$.

3 - <u>SOME REMARKS ABOUT</u> $\overline{w}^h(x)$

As C_{ip}^-, C_{ip}^+, b_i^h are convex combinations of the vertices of $S(x_i^h)$, using the linearity of w^h we have :

$$w^h(C_{ip}^-) + w^h(C_{ip}^+) = \sum_{j \in I_i^h} \gamma_{ij}\, w^h(x_j^h) \qquad \gamma_{ij} \geq 0, \sum_{j \in I_i^h} \gamma_{ij} = 2 \qquad (3.1)$$

I_i^h set of index such that $x_j^h \in S(x_i^h)$

$$w^h(b_i^h) = \sum_{j \in I_i^h} \gamma_j\, w^h(x_j^h) \qquad \gamma_j \geq 0\,, \sum_{j \in I_i^h} \gamma_j = 1. \qquad (3.2)$$

After (2.11), (3.1) and (3.2), we can rewrite $\mathcal{L}^h(u)w^h(x_i^h) + \ell(u, x_i^h) \geq 0$ as :

$$w^h(x_i^h) \leq \beta_i^h(x_i^h,u) \left[\frac{1}{h_i^2} \sum_{p=1}^{n} \lambda_p(x_i^h,u) \sum_{j\in I_i^h, j\neq i} \gamma_{pj}\, w^h(x_j^h) \right.$$

$$\left. + \frac{\| f(x_i^h) \|}{\| b_i^h - x_i^h \|} \sum_{j\in I_i^h} \gamma_j\, w^h(x_j^h) + \ell(u, x_i^h) \right] \tag{3.3}$$

$$\text{with } \beta_i^h(x_i^h,u) = \left[\sum_{p=1}^{n} \frac{(2-\gamma_{pi})}{h_i^2} \lambda_p(x_i^h,u) + \frac{\| f(x_i^h) \|}{\| b_i^h - x_i^h \|} + \alpha \right]^{-1} > 0.$$

Taking into account that all the factors that multiply $w^h(x_j^h)$ in the second member of (3.3) are non-negative we can easily prove (see [8]) :

THEOREM 1

There exists an unique $\overline{w}^h(x)$, maximum element of $\mathcal{W}^h$, i.e. $(\mathcal{P}^h)$ has an unique solution.

Furthermore the operator $\mathcal{L}^h$ verifies the following Discrete Maximum Principle (DMP) :

(DMP) : If C is a subset of vertices of Q^h satisfying $\mathcal{L}^h(u)\, w^h(x_i^h) \geq 0$,

$\forall x_i^h \in Q^h$, $\forall u \in U^h$, there exists Γ, $0 < \Gamma < 1$ such that : (3.4)

$$w^h(x_i^h) \leq \Gamma \left(\max_{x_i^h \in C} (w^h(x_i^h)) \vee 0 \right).$$

We can use this DMP to establish two important properties of $\overline{w}^h$.

The first one is that $\overline{w}^h$ is characterized by the fact that (3.3) becomes an equality for all $x_i^h \in Q^h$ for some $u \in U^h$ when we put $\overline{w}^h$ instead of w^h. This characterization allows us to compute $\overline{w}^h$ using iterative algorithms of the same type than those pesented in [8]. The value of u giving the equality will be used to define the optimal control $\hat{u}_h$.

The second one concerns the convergence of $\overline{w}^h$ to V. We have

THEOREM 2

The solutions $\overline{w}^h(x)$ of the approximate problems $(\mathcal{P}^h)$ converge uniformly to V(x), solution of $(\mathcal{P})$, i.e. :

$$\lim_{\|h\| \to 0} | V(x) - \overline{w}^h(x) | = 0, \forall x \in Q \qquad (3.5)$$

where $\|h\|$ is the maximum of the diameters of the simplex of Q^h. (see [8]).

The proof is achieved in two steps. We will briefly give here the main ideas.

In the first part we show

$$\underline{\lim}_{\|h\| \to 0} \quad \overline{w}^h \geq V. \qquad (3.6)$$

For that we regularize the elements of (1.8) by means of a convolution with a function of C^∞ $(\mathbb{R}^2)$ having a parameter $\rho > 0$. These functions w_ρ can be approximate by functions $w_{\rho,\alpha}$ with this property : the linear finite element $w^h_{\rho,\alpha}$, taking the same values of $w_{\rho,\alpha}$in the vertex of the triangulation Ω^h, belongs to $\mathcal{W}^h$. So,

$$\overline{w}^h \geq w^h_{\rho,\alpha} \qquad (3.7)$$

If we consider in (3.7) the lower limits for $\|h\| \to 0$, then the limits for $(\rho,\alpha) \to (0,0)$, we obtain

$$\underline{\lim}_{\|h\| \to 0} \quad \overline{w}^h \geq w. \qquad (3.8)$$

Finally, as w is an arbitrary element of $\mathcal{W}$, (3.6) is proved.

The second part is devoted to show

$$\overline{\lim}_{\|h\| \to 0} \quad \overline{w}^h \leq V. \qquad (3.9)$$

We consider a sequence of auxiliar problem $\mathcal{P}_n$ for which the controls u_n can take in (1.8) a finite number of values and the number of switchs within that set of values is, at most, n. If V_n is the solution of $\mathcal{P}_n$ we can show

$$V_1 \geqslant \ldots \geqslant V_n \geqslant V_{n+1} \geqslant V$$
$$\lim_{n\to\infty} V_n = V. \tag{3.10}$$

On the other hand we consider the discretized problem $\mathcal{P}^h_n$ for which we prove

$$\lim_{\|h\|\to 0} \overline{w}^h_n = V_n \tag{3.11}$$

$$\overline{w}^h_n \geqslant \overline{w}^h_{n+1} \geqslant \ldots \geqslant \overline{w}^h,\ \forall n. \tag{3.12}$$

So, $\overline{\lim\limits_{\|h\|\to 0}}\ \overline{w}^h \leqslant V_n$; then, using (3.10) we obtain (3.9). Finally (3.6) and (3.9) give (3.5).

4 - COMMENTS ON SOME APPLICATIONS

The idea of solving optimal control problems computing the maximum element of a suitable set of subsolutions of the Hamilton-Jacobi-Bellman equation has been recently applied to several problems. Remaining in the deterministic approach we have study in [9] the optimization of an electricity production system which comprise three hydraulic plants (two of pumped type) and seven thermic plants (one nuclear, two of coal, tow of fuel, one gas powered and one external). The numerical data have been provided by EDF (Electricity of France) : they describe a forecast of the French system for a week of the year 2000. Other application can be seen in [12] where several serial production/inventory systems are optimized.

Concerning the stochastic approach we can mention :

a) [11] devoted to the optimization of the system presented in [9] considering random perturbations in the demand ;

b) [7] in which the algorithm proposed in [10] for L(u) = Δ is used to obtain the optimal control of a bidimensional diffusion ;

c) [1] in which the numerical solution of an optimal correction problem for a damped random linear oscilator is studied.

First applications of the procedure just proposed in §2 and §3, as well as a comparison of these results with those obtained by other clasic methods [13], [14] and [17], will be presented in a special session of the next IEEE-CDC, Austin, 7-9 Dec. 1988.

5 - REFERENCES

[1] BANCORA M.C. - CHOW P. - MENALDI J.L., "*On the numerical approximation of an optimal correction problem*", submitted for publication.

[2] BENSOUSSAN A. - LIONS J.L., "*Applications des inéquations variationnelles en contrôle stochastique*", Dunod, Paris 1978.

[3] CIARLET P.G. - RAVIART P.A., "*Maximum principle and uniform convergence for the finite element method*", Computer Methods in Applied Mechanics and Engineering, 2 (1973), 17-31.

[4] CRANDALL M.G. - LIONS P.L., "*Viscosity solutions of Hamilton-Jacobi-Bellman equations*", Trans. AMS, 282, (1984), 487-502.

[5] FLEMING W.H. - RISHEL R., "*Optimal deterministic and stochastic control*", Springer-Verlag, Berlin, 1975.

[6] GONZALEZ R., "*Sur l'existence d'une solution maximale de l'équation de Hamilton-Jacobi-Bellman*", CRAS, Paris, 282, (1976),pp. 1287-1290.

[7] GONZALEZ R. - MEDINA M., "*Sobre la solución numérica del control óptimo de una difusión bidimensional*", ENIEF 87, Bariloche, July 1987.

[8] GONZALEZ R. - ROFMAN E., "*On deterministic control problems : an approximation procedure for the optimal cost*", Part I : The stationary case, SIAM J. on Control and Opt. 23, 2, (1985), 242-266 ; Part II : The non stationary case, SIAM J. on Control and Opt. 23, 2, (1985), 267-285.

[9] GONZALEZ R. - ROFMAN E., "*On the optimization of a short-run model of energy production systems*", Lecture Notes in Control and Inf. Sci., Proceedings of the 12th IFIP Conf. Budapest (1985), Springer-Verlag 1986, 757-765.

[10] GONZALEZ R. - ROFMAN E., "*On stochastic control problems. An algorithm for the value function and the optimal policy*", 13th IFIP Conference on System Modelling and Optimization, Tokyo, Aug. 31 -Sept. 4 1987. To appear on Springer Verlag, Lect. Notes in Control and Inf. Sc.

[11] GONZALEZ R. - ROFMAN E., "*On the computation of optimal control policies of energy production systems with random perturbations*", Proc. 26th CDC-IEEE Los Angeles, Dec. 1987, pp. 312-313.

[12] KABBAJ F. - MENALDI J.L. - ROFMAN E., "*Variational approach of serial multi-level production/inventory systems*", RR INRIA 692, Juin 1987.

[13] KUSHNER H., "*Probability methods for approximations in stochastic control and for elliptic equations*", Academic Press, New York (1977).

[14] LIONS P.L. - MERCIER B., "*Approximation numérique des équations de Hamilton-Jacobi-Bellman*" RAIRO Analyse Numérique, 14, (1980), 369-393.

[15] LIONS P.L. - MENALDI J.L., "*Optimal control of stochastic integrals and Hamilton-Jacobi-Bellman equations*", II SIAM J. on Control and Opt. 20, 1, January 1982.

[16] MENALDI J.L., "*Sur les problèmes de temps d'arrêt, contrôle impulsionnel et continu correspondant à des opérateurs dégénérés*", Th. d'Etat, Université Paris Dauphine, Dec. 1980.

[17] THEOSYS, "*Commande optimale de systèmes stochastiques*", RAIRO Automatic 18, (1984), pp. 225-250.

A HIERARCHICAL BARGAINING MODEL IN ENERGY MANAGEMENT

List of authors: J. Ruusunen, H. Ehtamo, and R.P. Hämäläinen

Address:
Systems Analysis Laboratory
Helsinki University of Technology
Otakaari 1M
02150 Espoo, Finland

Keywords: Energy management, dynamic games, bargaining

The purpose of an electric power pool is to reduce the cost of generating electricity by transferring electric power between the power plants that are controlled by individual decision makers. During high local demand a system can transfer energy from the network into the system and thus achieve cost savings. At the same time other systems produce electricity into the network such that power balance in the network is met. The benefits of receiving energy in one period are then compensated by an energy transfer into the network in some other period. As a whole, the the pool can thus achieve cost savings.

The problem of equitable sharing of the benefits of cooperation during the planning horizon is a bargaining problem in the dynamic framework. We shall formulate the energy bargaining model in the dynamic framework and propose a new way of dividing the cost savings within the power pool. The energy exchange contract is determined on the basis of the Nash bargaining scheme. In our previous studies we have presented a hierachical approach to solve Nash bargaining problems in the dynamic framework. This approach is extremely convenient in this application.

SPECULATIONS ON POSSIBLE DIRECTIONS AND APPLICATIONS FOR THE DECOMPOSITION METHOD

G. Adomian
Department of Mathematics
The University of Georgia
Athens, Georgia 30602

The decomposition method has now solved very accurately a rather wide class of nonlinear differential and partial differential equations [1] showing some significant advantages over other methods. Once a problem is modeled with a specific equation (linear, nonlinear, deterministic, stochastic, ordinary or partial differential equation) with physically correct given conditions, the method solves the equation without linearizations, perturbations, closure approximations, white noise assumptions, or discretization. Certainly, much remains to be done on the theoretical foundation and the precise limitations. Rather than a drawback, this is a fascinating challenge for further work which is beginning to be borne out by the increasing work in this field particularly in Torino by Professor N. Bellomo [2] and his co-workers [2] as well as by many others. The range of problems solved and the rather remarkable accuracy obtained - the fact that nonlinear systems with stochastic parameters can be solved and the fact that the work has applied effectively to parabolic, elliptic, and hyperbolic equations - certainly suggest this is a useful and very computational method for frontier applications. Proof of convergence and convergence rate, error estimates, and perhaps better generation of Adomian's A_n polynomials are fertile areas for further study and dissertations. Many other research topics are in the area of applications; some are discused in [3].

Let us point out some speculations on some interesting possible future applications pointing out that some of these applications require the development of a correct mathematical model before decomposition can possibly solve them. It is not useful to apply the method to many existing models since they have already been linearized and otherwise simplified for mathematical tractability. Thus it is up to the expert in physics, engineering, biology, economics, agriculture, etc., to model the problems retaining the nonlinearities, stochasticity, delays, etc., since the physically correct solution can be very different from that ^btained from the simplified models. Also since the technique does not require discretization, it is evident that substantially less computing time may be involved in a difficult problem such as Navier-Stokes equations [4].

Nevertheless, some possible applications which represent an exciting challenge are areas such as nonlinear and possibly stochastic and multidimensional optiomal control theory, hypersonic flow, quantum theory and gravitation, generalization of the Kalman filter, and problems of large space structures such as vibration, heating, etc., [3].

Before going into these areas, let's look briefly at some illustrative decomposition examples chosen for clarifying procedure rather than for difficulty.

Consider an ordinary differential equation $d^2u/dx^2 - 40xu = 2$, $u(-1) = u(1) = 0$. Let $L = d^2/dx^2$ and write [1]

$$Lu = 2 + 40xu$$

$$u = c_1 + c_2x + L^{-1}(2) + L^{-1}(40xu)$$

Let $u_0 = c_1 + c_2x + L^{-1}(2) = c_1 + c_2x + x^2$ and let $u = \sum_{n=0}^{\infty} u_n$.

The components of u are given by

$$u_{n+1} = L^{-1}40xu_n$$

for $n \geq 0$ thus

$$u_1 = L^{-1}40xu_0 = (20/3)c_1x^3 + (10/3)c_2x^4 + 2x^5$$

Similarly

$$u_2 = (80/9)c_1x^6 + (200/63)c_2x^7 + (10/7)x^8$$

We continue to some n-term approximation $\varphi_n = \sum_{i=0}^{n-1} u_i$ which approaches $u = \sum_{n=0}^{\infty} u_n$ as $n \to \infty$ [1]. If we write φ_3 as an approximation,

$$\begin{aligned}\varphi_3 &= u_0 + u_1 + u_2 \\ &= c_1 + c_2x + x^2 + (20/3)c_1x^3 + (10/3)c_2x^4 \\ &\quad + 2x^5 + (80/9)c_1x^6 + (200/63)c_2x^7 + (10/7)x^8\end{aligned}$$

Imposing the boundary conditions at -1,1, we write $\varphi_3(1) = \varphi_3(-1) = 0$ from which we get

$$\begin{bmatrix} 149/9 & 473/63 \\ 29/9 & -53/63 \end{bmatrix} \cdot \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} -31/7 \\ -3/7 \end{bmatrix}$$

from which c_1, c_2 are evaluated. Substituting φ_n onto the left side of the differential equation, we should get the right side, or 2, if the approximation is sufficient. We note that the 12-term approximation yields 2.000000 or seven-digit accuracy.

On R^3 with $L_x = \partial^2/\partial x^2$, $L_y = \partial^2/\partial y^2$, $L_z = \partial^2/\partial z^2$ we write

$$[L_x + L_y + L_z]u = f(x,y,z) + k(x,y,z)u$$

Solve for each linear operator in turn. Operate on each of the three equations with the appropriate inverse and write

$$u = \varphi_x + L_x^{-1}f - L_x^{-1}ku - L_x^{-1}(L_y + L_z)u$$

$$u = \varphi_y + L_y^{-1}f - L_y^{-1}ku - L_y^{-1}(L_z + L_x)u$$

$$u = \varphi_z + L_z^{-1}f - L_z^{-1}ku - L_z^{-1}(L_x + L_y)u$$

where $\varphi_x, \varphi_y, \varphi_z$ are the homogeneous solutions. Adding and dividing by 3.

$$u = u_0 + Ku$$

with

$$u_0 = (1/3)(\varphi_x + \varphi_y + \varphi_z + (L_x^{-1} + L_y^{-1} + L_z^{-1})f)$$

$$K = (1/3)(L_x^{-1} + L_y^{-1} + L_z^{-1})k + L_x^{-1}(L_y + L_z))$$

$$+ L_y^{-1}(L_z + L_x) + L_z^{-1}(L_x + L_y))$$

assuming $u = \sum_{n=0}^{\infty} u_n$,

$$u_{n+1} = K\,u_n$$

so all components are determined. The inverse operators are double integrations leading to two constants of integration to be determined by forcing u_n to satisfy the given condition.

Suppose $k = k(u)$ so the equation becomes nonlinear. The nonlinear term is expanded as $\sum_{n=0}^{\infty} A_n$ where the A_n are Adomian polynomials [1,3] generated for the nonlinear term and the procedure is as before except that the u_{n+1} will involve an A_n term. Since each A_n depends only on $u_0, u_1, \ldots . u_n$, the solution can be obtained essentially as easily as in the linear case.

Rather than further discussion of the methodology on which there is now a considerable published literature in the U.S. and Europe, let us speculate on some applications which appear to be possible in the very near future although they require the modelling expertise of theorists concerned primarily with each of those areas.

Some of these, in the authors opinion, are

1) optimal control for nonlinear and, stochastic, and even multidimensional systems,
2) hypersonic flow, turbulence, single-stage-to-orbit flight essential for shuttles which can be used for the construction of space stations,
3) quantum theory and gravitation, and
4) generalizations of Kalman filtering.

Because of page and time limitations we discuss only the first two here.

1) Suppose we consider a nonlinear, possibly stochastic or even multidimensional systems which we want to control in some optimal way. For a linear control system with a quadratic performance index, of course an analytical solution can be made. Consider the state equations

$$\dot{x}(t) = f(x_1,\ldots,x_n;\ u_1,\ldots,u_m;t)$$

is, a set of n nonlinear differential equations with $x(t)$ representing a state vector with n components $f_1,\ldots,f_n$, and $x(t_0)$ a given initial vector. Define, for example [5] a performance functional $J(x,u,t)$ given by

$$J = \phi[x(t_1),\ t_1] + \int_{t_0}^{t_1} F(x,u,t)\ dt$$

where ϕ and F are scalar functions with necessary smoothness properties. Let $p = [p_1,\ldots,p_n]^T$ be a vector of Lagrange multipliers and form an augmented functional

$$J' = \phi[x(t_1),t_1] + \int_{t_0}^{t_1} [F(x,u,t) + p^T\ (f-\dot{x})]\ dt$$

Integration by parts leads to

$$J' = \varphi - [p^T x] \Big|_{t_0}^{t_1} + \int_{t_0}^{t_1} [H + \dot{p}^T x] \, dt$$

with H defined as

$$H(x,u,t) = F(x,u,t) + p^T f$$

if u is defined on $t_0 \le t \le t_1$, we vary u and find the variation $\delta J'$ corresponding to δu, leading to the n adjoint equations,

$$\dot{p}_i = - \frac{\partial H}{\partial x_i}$$

so we have a system of $2n$ nonlinear differential equations with two-point boundary conditions. Although this approach has been discussed by R.E. Bellman and many others perhaps most recently in [5], analytical solution has usually not been possible except by numerical methods. We now have a promising and potentially valuable alternative since such systems of nonlinear differential equations have been solved (even for the stochastic and/or multidimensional cases) in a analytic approximation by the decomposition method [1-3].

Another possibility is through solution by decomposition of the matrix Riccati equation which appears in invariant imbedding and neutron transport theory as well as modern control theory. Consider

$$R'(x) = B(x) + D(x)R(x) + R(x)D(x) + R(x)B(x)R(x)$$
$$R(0) = 0$$

where B, D, R are continuous $n \times n$ non-negative matrices. Suppressing the argument x, we have

$$R' = B + DR + RD + RBR$$

If $L = d/dx$

$$LR = B + HR + NR$$

where $LR = R'$, $HR = DR + RD$, and NR represents a nonlinear operator on R . Since $R(0) = 0$, operation with L^{-1} on both sides yields

$$R = L^{-1}B + L^{-1}HR + L^{-1}NR \ .$$

Let R and NR be written in terms of Adomian's A_n polynomials. For R this is equivalent to writing $R = \sum_{n=0}^{\infty} R_n$. For NR we write $\sum_{n=0}^{\infty} A_n$. Identify $R_0 = L^{-1}B$ then

$$R_0 = L^{-1} B$$

$$R_1 = L^{-1} H R_0 + L^{-1} A_0$$

$$R_2 = L^{-1} H R_1 + L^{-1} A_1$$

.

.

.

$$R_n = L^{-1} H R_{n-1} + L^{-1} A_{n-1}$$

for $n \geq 1$. The A_n for NR are given by [1]

$$A_0 = R_0 B R_0$$

$$A_1 = R_0 B R_1 + R_1 B R_0$$

$$A_2 = R_1 B R_1 + R_0 B R_2 + R_2 B R_0$$

$$A_3 = R_0 B R_3 + R_3 B R_0 + R_1 B R_2 + R_2 B R_1$$

$$A_4 = R_2 \; B \; R_2 + R_0 \; B \; R_4 + R_4 \; B \; R_0$$
$$+ R_1 \; B \; R_3 + R_3 \; B \; R_1$$
$$\vdots$$

so that

$$R_0 = L^{-1} \; B$$

$$R_1 = L^{-1} \; H \; R_0 + L^{-1} \; R_0 \; B \; R_0$$

$$R_2 = L^{-1} \; H \; R_1 + L^{-1} \; (R_0 \; B \; R_1 + R_1 \; B \; R_0)$$

$$R_3 = L^{-1} \; H \; R_2 + L^{-1} \; (R_1 \; B \; R_1 + R_0 \; B \; R_2 + R_2 \; B \; R_0)$$

$$\vdots$$

Finally since $HR = DR + RD$

$$R_0 = L^{-1} \; B$$

$$R_1 = L^{-1}(D \; R_0 + R_0 \; D) + L^{-1}(R_0 \; B \; R_0)$$

$$R_2 = L^{-1}(DR_1 + R_1 \; D) + L^{-1}(R_0 \; B \; R_1 + R_1 \; B \; R_0)$$

$$R_3 + L^{-1}(D \; R_2 + R_2 \; D) + L^{-1}(R_1 \; B \; R_1 + R_0 \; B \; R_2 + R_2 \; B \; R_0)$$

$$\vdots$$

An n-term approximant is $\phi_n = \sum_{i=0}^{n-1} R_n$ which approaches $R = \sum_{i=0}^{\infty} R_n$ as $n \to \infty$. Thus given B, D, a specific R can be calculated to a desired approximation. Accuracy has been demonstrated in [6].

(2) Hypersonic Flow: The present approach to hypersonic flow problems is computational fluid dynamics (CFD), and intensive work is being done to develop appropriate CFD computer programs for the hypersonic case. With continuing rapid developments in supercomputers, this emphasis is certainly appropriate. Yet, another methodology now appears promising which is quite different and seems to have a high potential for important advantages as well as a probably high adaptability to supercomputers. This is the decomposition method.

It yields a rapidly converging series solution in analytic form. It requires no linearization, perturbation, closure approximations, or assumption of special mathematically tractable stochastic processes such as delta-correlated processes. Probably most important is the fact that discretization into grids is unnecessary. Hence, computation should be enourmously less, and the difficulty of different time scales in turbulence is avoided.

In the types of fluid flow which interest us, velocity, density, and pressure are stochastic, not constants. Present treatment of Navier-Stokes equations solves a simplistic model, not real behavior. Turbulence is a strongly nonlinear, strongly stochastic phenomenon and cannot be understood by linearized perturbative treatments. The theories of physics are perturbative theories and the theories of mathematics are for linear operators (other than some ad hoc methods for special nonlinear equations). What is needed is a way of solving one or more nonlinear stochastic operator equations whether algebraic, differential, delay-differential, partial-differential, or systems of such equations. The computational accuracy of a supercomputer is dependent on the sophistication of the mathematical methods programmed into it. Typical calculations consider millions of discrete time intervals mad small enough so trajectories between them can be taken as low-order polynomials, e.g., quadratics. If stochasticity is involved, then Monte Carlo methods are used which inserts randomness but not the properly correlated randomness which is present in the physical problem.

In generalized hydrodynamics, the form of Navier-Stokes equations is kept, but time and distance scales are introduced so one can go beyond continuum approximation and take account of molecular structure. However, application to a real situation becomes simply a test of the validity of the linear approximations, as pointed out in the literature. Fluctuations are , as usual, assumed "small," and delayed effects, due to the fact that responses cannot be instantaneous, are ignored.

When one studies airflow about aircraft surfaces, computations are made tens of millions of points, and it is felt that increasing the volume of computation to the limit in an ultimate extrapolation, supercomputers will yield complete accuracy. Not only does this ignore stochasticity, it ignores the sensitivity of nonlinear stochastic systems to very slight changes in the model - in fact, to changes essentially undeterminable by measurement.

To solve an aircraft problem on contemplated next-generation computers, a 3-dimensional mesh is generated which discretizes the system of nonlinear partial differential equations into a million, a hundred million, or perhaps a billion coupled difference equations in as many unknowns. One begins to see then the tremendous data handling problem, the necessity for improved algorithms, and the need for still greater computational speed. We may also have many unknowns at each point, and, as we have pointed out, the system nonlinearities and random fluctuations need to be taken into consideration. Since usually solutions are iterative - first solving an approximation to the original system of differential equations and then improving the solution by repeated substitution of each new solution - parallel processing is complicated by the difficulty of partitioning the work so each processor can work independantly. This is being pursued by many ingenious ideas necessitated by the brute force method of discretization.

In all such problems we need to be able to solve coupled systems of nonlinear (and generally stochastic as well) partial differential equations with complex boundary conditions and possible delayed effects. These systems are linearized and discretized (and the stochastic aspects either ignored or improperly dealt with) so the various numerical approximation methods can be used. This requires faster and faster supercomputers to do these computations in a reasonable time.

Unfortunately the further developments in supercomputers can quite possibly give wrong answers because even a single one-dimensional nonlinear differential equation without stochasticity in coefficients, inputs, and boundary conditions - let alone vector partial differential equations in space and time with nonlinear and/or stochastic parameters - are not solved exactly. Real systems are nonlinear and stochastic. When you throw out these "complications," you have a different problem! When you linearize and use perturbative methods, you solve a mathematized problem, not the physical problem. The model equations, even before the linearization, discretization, etc. are already wrong because the stochastic behavior is generally not incorporated or is incorporated incorrectly as an afterthought.

Our approach to hypersonics, using decomposition, will be based on previous work on Navier-Stokes [3,4] which showed an analytic solution can be carried out. For hypersonic cases, additional effects are present changing the model equations but the approach is similar. Discussion of a rather global mathematical methodology, let alone the huge subject of hypersonics and turbulence is, of course, not addressable here. We can only call attention now to the possibility of some promising alternatives to the present approaches [3].

1. Adomian, G. Nonlinear Stochastic Operator Equations, Academic Press 1986.

2. Bellomo, N. and Riganti, R. Nonlinear Stochastic Systems in Physics and Mechanics, World Scientific Publ. Co., 1987.

3. Adomian, G. Application of Nonlinear Stochastic Systems Theory to Physics, Reidel, 1988.

4. Adomian, G. "Solution of Navier-Stokes Equations," Comp. & Math. with Applic., 12A, no. 11, 1986, pp. 1119-1124.

5. Barnett, S. and Cameron, R.G. Introduction to Mathematical Control Theory, Clarendon Press 1985.

6. Adomian, G., Pandolfi, M., and Rach, R., An Application of the Matrix Riccati Equation to a Neutron Transport Process, J. Math. Anal. and Applic., in publication.

Fuzzy arithmetic in qualitative reasoning

Didier DUBOIS and Henri PRADE

Laboratoire Langages et Systèmes Informatiques
Université Paul Sabatier, 118 route de Narbonne
31062 TOULOUSE Cédex (FRANCE)

The paper provides a preliminary exploration of the application of fuzzy arithmetic and fuzzy approximate reasoning techniques to qualitative reasoning problems considered in Artificial Intelligence. More specifically, this investigation is done along three lines : constraint propagation with ill-known values, handling of orders of magnitude in terms of fuzzy intervals or by means of fuzzy relations.

1 - Introduction

Reasoning about the behavior of systems in a qualitative way is interesting in two kinds of circumstances : i) when the system under consideration is complex and the data available about it are pervaded with imprecision or even vagueness ; ii) when it is sufficient to have a qualitative view of the system and of its behavior, and this qualitative view is not only easier to get than a more precise one from a computational point of view, but also easier to understand. From the beginning of the eighties there have been a growing interest about qualitative reasoning in Artificial Intelligence ; see (Bobrow, 1984 ; Dormoy, 1987) for an introduction. The intended purpose of this research is mainly to provide understandable explanations of the behavior of complex systems from their qualitative description. The modeling is done in terms of variables which are potentially real-valued, but the analysis and the description of the system behavior is made only in terms of three values usually, namely "-", "0" and "+", corresponding to whether the variables are negative, zero or positive. Independently, works motivated by research in qualitative economics, have been developed about qualitative controllability and observability of linear dynamical systems where real-valued variables are approximated in terms of the same three values ; see Travé and Kaszkurewicz (1986) for instance.

From the end of the seventies, fuzzy set and possibility theory (Zadeh, 1978 ; Dubois and Prade, 1985), whose introduction was initially motivated by the modeling of complex and ill-known systems, has been considerably developed both from a theoretical and an applied point of view in various directions ; particularly, fuzzy arithmetic (Dubois and Prade, 1980, 1987) enables us to handle ill-known quantities in an easy way which generalizes interval analysis, and besides a methodology for approximate reasoning (Bellman and Zadeh, 1977) has been settled in the fuzzy set framework. Until now there have been no serious attempt to use fuzzy techniques in qualitative reasoning problems in Artificial Intelligence --if we except some hints (Raiman, 1985) and preliminary works (d'Ambrosio, 1987)-- although it would be desirable in some cases to have a finer and less sharp description of the values of the variables than the one provided by "-", "0" or "+". Particularly, the sign of the difference between two positive quantities cannot be determined without any information about their respective order of magnitude.

This paper investigates what may be the use of fuzzy arithmetic and fuzzy set-based approximate reasoning techniques in qualitative reasoning problems. First, a general approach for refining interval values attached to variables by exploiting constraints which must be satisfied by these variables, is extended to fuzzy set values. Then, a fuzzy interval-based approach is proposed for handling orders of magnitude in arithmetic operations and a valid approximation technique is used in order to insure a closure property of the operations restricted to the considered fuzzy values. The interest of fuzzy intervals for interfacing symbolic information and numerical data, is emphasized. Then another way of dealing with orders of magnitude based on approximate equality relations is investigated. The concluding remarks point out some other contributions of fuzzy logic to qualitative control and to qualitative descriptions of systems behavior.

2 - Constraint propagation with fuzzy values

2.1 - General discussion

Let $X_1, \ldots, X_n$ denote single-valued real variables. Let A_i be a subset of the real line which is known to restrict the possible values of X_i, and let R be a relation which must be satisfied by the X_i's and which acts as a constraint on $(X_1, \ldots, X_n)$. Then, the refinement of the possible ranges of the variables X_i's taking into account R, leads to update the possible range of each variable X_i into a new subset A'_i in the following way

$$A'_i = \{x_i \in A_i \mid \exists\, x_j \in A_j, j = 1,n, j \neq i \text{ and } (x_1, \ldots, x_i, \ldots, x_n) \in R\} \qquad (1)$$

More generally in case of several constraints represented by relations R_k, $k = 1,r$, we can iterate this refinement procedure on each variable taking successively each relation into account over and over until no more changes occur in the updated ranges. This is known in Artificial Intelligence as the Waltz algorithm ; see Davis (1987) for a detail study of this procedure both from an implementation and an application point of view. Let us consider a simple example. Let $n = 3$, $A_1 = [0,2]$, $A_2 = [1,3]$ and $A_3 = [0,2]$ and the constraint $X_1 + X_2 = X_3$. Then we get $A'_1 = [0,1]$, $A'_2 = [1,2]$ and $A'_3 = [1,2]$. Observe that any triple of values in the Cartesian product $A'_1 \times A'_2 \times A'_3$ is not necessarily feasible, e.g. $\nexists\, x_3 \in A'_3$ such that $x_1 + x_2 = x_3$ with $x_1 = 1$ and $x_2 = 2$.

The definition (1) expresses that A'_i is obtained as the intersection of A_i with the result of the composition of the relation R with the Cartesian product of the A_j's except A_i. This can be readily extended to the case where the A_i's are fuzzy sets and/or R represents a fuzzy constraint ; i.e.

$$\forall i, \forall x_i,\ \mu_{A'_i}(x_i) = \min[\mu_{A_i}(x_i), \quad \sup_{\substack{x_j \\ j=1,n\,;\, j\neq i}} \quad \min(\mu_R(x_1, \ldots, x_n), \quad \min_{j=1,n\,;\, j\neq i} \quad \mu_{A_j}(x_j))] \qquad (2)$$

where μ denotes the membership functions (whose range are [0,1]) of the corresponding fuzzy sets and relation. When R is an ordinary relation such that X_i is a function f of the other variables X_j, A'_i is a fuzzy set which can be obtained by applying f, in the sense of fuzzy set and possibility theory, to the A_j's ($j \neq i$), i.e.

$$\forall x_i,\ \mu_{A'_i}(x_i) = \min[\mu_{A_i}(x_i), \sup_{f(x_j,\, j=1,n,\, j\neq i) = x_i} \ \min_{j=1,n\,;\, j\neq i} \mu_{A_j}(x_j)] \qquad (3)$$

When the A_j's are fuzzy intervals and f is monotonic with respect to each variable and can be expressed in terms of arithmetic operations, the A'_i's are fuzzy intervals which can be easily computed using results of fuzzy arithmetic ; see Dubois and Prade (1985, 1987). This extends the fact that, for instance, in the above example the A'_i's can be obtained as the result of operations on intervals ; namely $A'_1 = A_1 \cap (A_3 \ominus A_2)$, $A'_2 = A_2 \cap (A_3 \ominus A_1)$, $A'_3 = A_3 \cap (A_1 \oplus A_2)$, where the circled symbols are used for denoting the extension of arithmetic operations to intervals. Indeed fuzzy arithmetic generalizes interval arithmetic. Note that the refinement is obtained in (2) in one step, in the sense that refined A'_j's cannot enable us to obtain a more restrictive A'_i. This can be easily checked ; indeed, taking n = 2 for notational convenience, we have

$$\min(\mu_{A_1}(x_1), \sup_{x_2} \min(\mu_R(x_1,x_2), \mu_{A'_2}(x_2)))$$

$$= \sup_{x_2} \min(\mu_{A_1}(x_1), \mu_R(x_1,x_2), \sup_{x_1} \min(\mu_{A_1}(x_1), \mu_R(x_1,x_2)), \mu_{A_2}(x_2))$$

$$= \mu_{A'_1}(x_1) \text{ since obviously } \min(\mu_{A_1}(x_1), \mu_R(x_1,x_2)) \leq \sup_{x_1} \min(\mu_{A_1}(x_1), \mu_R(x_1,x_2))$$

In fact, (2) can be viewed as a particular case of the general approach to approximate reasoning initiated in Bellman and Zadeh (1977) and developed in Zadeh (1979), namely, all the pieces of information are conjunctively combined and then the result is projected on the domain of the variable(s) in which we are interested. Indeed (2) can be equivalently rewritten

$$\forall i,\ \forall x_i,\ \mu_{A'_i}(x_i) = \sup_{\substack{x_j \\ j=1,n\,;\, j\neq i}} \min(\mu_R(x_1, \ldots, x_n), \mu_{A_1}(x_1), \ldots, \mu_{A_i}(x_i), \ldots, \mu_{A_n}(x_n)) \qquad (4)$$

In case of several relations R_k the combination/projection method leads to the following updating scheme where the R_k's are replaced by their cylindrical extensions when they do not involve all the variables

$$\forall i, \forall x_i,\ \mu_{A'_i}(x_i) = \sup_{\substack{x_j \\ j=1,n\,;\, j\neq i}} \min(\min_{k=1,r} \mu_{R_k}(x_1, \ldots, x_n),\ \min_{j=1,n} \mu_{A_j}(x_j)) \qquad (5)$$

$$\leq \min_{k=1,r} [\min(\mu_{A_i}(x_i), \sup_{\substack{x_j \\ j=1,n\,;\, j\neq i}} \min(\mu_{R_k}(x_1, \ldots, x_n), \min_{j=1,n\,;\, j\neq i} \mu_{A_j}(x_j)))] \qquad (6)$$

The inequality (6) expresses that if we take into account each R_k separately in the refinement process, we are not sure, even if we iterate the procedure as in the Waltz algorithm, of obtaining the most accurate refinement for each variable range. However, what is got by (6) is obviously valid and more easy to compute in general.

Note that in case of binary relations, the Waltz procedure (i.e. the separate processing of the R_k's) yields the most accurate result given by (5), provided there is at most one relation R_k between any pair of variables (x_i,x_j) and that there is no cycle in the non-oriented graph whose nodes correspond to the variables and edges to the binary relations. Indeed, for instance with $n = 3$ and two relations, we have

$$\begin{aligned}\mu_{A'_1}(x_1) &= \min(\mu_{A_1}(x_1), \sup_{x_2,x_3} \min(\mu_R(x_1,x_2), \mu_{R'}(x_2,x_3), \mu_{A_2}(x_2), \mu_{A_3}(x_3)) \\ &= \min(\mu_{A_1}(x_1), \sup_{x_2} \min(\mu_R(x_1,x_2), \min(\mu_{A_2}(x_2), \sup_{x_3} \min(\mu_{R'}(x_2,x_3), \mu_{A_3}(x_3)))))\end{aligned} \quad (7)$$

2.2 - *Fuzzy equalities and inequalities*

In this subsection, we consider particular fuzzy relations which are of interest in practice for qualitative reasoning. Approximate equalities or strong inequalities (e.g. 'much greater than") are examples of binary fuzzy relations which can be easily handled using fuzzy arithmetic techniques. Indeed an approximate equality can be modelled by a fuzzy relation E of the form $\mu_E(x,y) = \mu_L(|x - y|)$, for instance

$$\forall x, \forall y,\ \mu_E(x,y) = \max(0, \min(1, \frac{\delta + \varepsilon - |x - y|}{\varepsilon})) = \begin{cases} 1 \text{ if } |x - y| \le \delta \\ 0 \text{ if } |x - y| \ge \delta + \varepsilon \\ \dfrac{\delta + \varepsilon - |x - y|}{\varepsilon} \text{ otherwise} \end{cases} \quad (8)$$

where δ and ε are respectively positive and strictly positive parameters which modulate the approximate equality. Then the approximate equality of variables X and Y (in the sense of E) will be written under the form of the equality

$$X - Y = L \quad (9)$$

with the following intended meaning : the possible values of the difference X - Y are restricted by the fuzzy set L. Here L is a fuzzy interval centered in 0, i.e. $L = -L$ since $\mu_L(d) = \mu_L(-d)$ or if we prefer $\mu_E(x,y) = \mu_E(y,x)$. Similarly a strong inequality can be modelled by a relation I of the form $\mu_I(x,y) = \mu_K(x - y)$, for instance

$$\forall x, \forall y,\ \mu_I(x,y) = \max(0, \min(1, \frac{x - y - \lambda}{\rho})) = \begin{cases} 1 \text{ if } x \ge y + \lambda + \rho \\ 0 \text{ if } x \le y + \lambda \\ \dfrac{x - y - \lambda}{\rho} \text{ otherwise} \end{cases} \quad (10)$$

where $\lambda \ge 0$ and $\rho > 0$. The constraint 'X is much greater than Y' (in the sense of I) can then be written

$$X - Y = K \quad (11)$$

where K is a fuzzy interval such that $K = [K,+\infty)$ (with $\mu_{[K,+\infty)}(t) = \sup_{s \le t} \mu_K(s)$), i.e. K identifies itself as the set of values equal or greater than a value restricted by K.

If we know for instance that 'X$_1$ is approximately equal to X$_2$' (i.e. $X_1 - X_2 = L$) and that 'X$_2$ is much greater than X$_3$' (i.e. $X_2 - X_3 = K$), we can deduce that

$$X_1 - X_3 = L \oplus K$$

where $\oplus$ denotes the addition extended to fuzzy intervals[1] (see Dubois and Prade (1980, 1987)). It can

[1] Let $\odot$ denote the extension of an arithmetic operation ^ to fuzzy sets of the real line. $\odot$ is defined by $\mu_{K \odot L}(u) = \sup_{s,t,\ u=s \wedge t} \min(\mu_K(s), \mu_L(t))$. Besides $\mu_{f(K)}(t) = \sup_{t=f(s)} \mu_K(s)$. When ^ is the addition and K and L are trapezoids represented by the abscissas of the endpoints of their parallel sides, it can be proved that $(k_1, k_2, k_3, k_4) \oplus (l_1, l_2, l_3, l_4) = (k_1 + l_1, k_2 + l_2, k_3 + l_3, k_4 + l_4)$ (k_i or l_j may be equal to $-\infty$ or $+\infty$).

be proved that it means that it is certain that $X_1 \geq X_3 + \lambda - (\delta + \varepsilon)$ and that the value of the difference $X_1 - X_3$ belongs to $L \oplus K$ at the degree 1 as soon as $X_1 \geq X_3 + \lambda + \rho - \delta$. See Figure 1. Then depending on the respective values of the parameters, X_1 is still greater than X_3 (but may be not as much as X_2 with respect to X_3) (if $\lambda > \delta + \varepsilon$), or we are only sure that X_1 is not much smaller than X_3 (if $\lambda + \rho < \delta$). Moreover, if we know that $X_3 = A_3$, we shall get

$$X_1 = A'_1 = A_3 \oplus L \oplus K$$

This is a particular case of (7) where $R = E$, $R' = I$, $A_2 = (-\infty,+\infty) = A_1$.

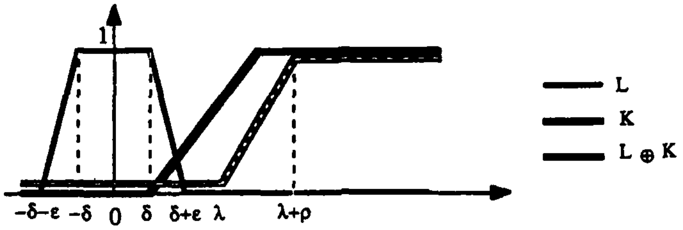


Figure 1

2.3 - Linear constraints

Another worth-considering particular case of the general problem presented in 2.1 is the one of linear systems of constraints. For sake of simplicity, we only briefly discuss linear systems with two variables and two constraints of the form

$$\begin{cases} a_1X_1 + b_1X_2 = A_3 \\ a_2X_1 + b_2X_2 = A_4 \end{cases}$$

where A_3 and A_4 are fuzzy sets of real numbers, and the other coefficients are real numbers. Note that each of these constraints implicitly defines a fuzzy relation which restricts the possible values of the pair (X_1,X_2). Provided that $a_1b_2 - a_2b_1 \neq 0$, we can deduce, using (3), that

$$X_1 = A'_1 = \frac{b_2A_3 \ominus b_1A_4}{a_1b_2 - a_2b_1} \; ; \; X_2 = A'_2 = \frac{a_2A_3 \ominus a_1A_4}{a_2b_1 - a_1b_2} \qquad (12)$$

with $A_1 = A_2 = (-\infty,+\infty)$; *see the footnote 1 for the definition of the extended difference* $\ominus$ and of the product of a fuzzy quantity by a scalar. If the constraints are changed into $a_1X_1 + b_1X_2 = X_3$ and $a_2X_2 + b_2X_2 = X_4$, with $X_3 = A_3$ and $X_4 = A_4$, the ranges of possible values of X_3 and X_4 are respectively updated into $A'_3 = A_3 \cap (a_1A'_1 \oplus b_1A'_2)$ and into $A'_4 = A_4 \cap (a_2A'_1 \oplus b_2A'_2)$.

More generally, the coefficients in linear systems may be ill-known. Then direct extensions of (12) can still be used where the a_i's and b_j's are replaced by fuzzy quantities and where we use the product and the quotient defined in fuzzy arithmetics. However in that case we get ranges which are still valid but may be larger than the actual ranges. This is due to the interactivity constraint which requires that the values of a_i or b_j should be the same at the numerators and the denominators in (12), even if the coefficients are *ill-known*, and which is forgotten in a straightforward calculation. This interactivity constraint should be taken into account for obtaining the actual ranges. See Dubois (1987) for a general discussion of fuzzy linear programming.

3 - Fuzzy intervals and orders of magnitude

Standard qualitative reasoning distinguishes between values which are strictly negative (-), zero (0) or strictly positive (+), and is based on the exploitation of the following tables for the addition and the product

⊕	0	+	-	?
0	0	+	-	?
+	+	+	?	?
-	-	?	-	?
?	?	?	?	?

⊗	+	-	?	0
+	+	-	?	0
-	-	+	?	0
?	?	?	?	0
0	0	0	0	0

Tables 1

where ? denotes the completely unknown value corresponding to the range $(-\infty,+\infty)$. However, if we know for instance that $X_1 = + \; ; \; X_3 = + \; ; \; X_1 + X_2 = X_3$
we can only deduce $X_2 = ?$ (while if $X_1 = 0$, we get $X_2 = +$). Another simple example of the undesirably limited representation power of the above calculus is the following

$$\text{if } X_1 = + \text{ and } X_2 = + \text{ then } X_3 = X_1 + X_2 = +$$

then the fact that $X_3 > X_1$ and $X_3 > X_2$ is forgotten. These kinds of ambiguities could be removed, if a more precise knowledge about the orders of magnitude, which is often available, could be modelled. Indeed we have in the general case for the first above example

$$X_1 = A_1 \; ; \; X_2 = A_2 \; ; \; X_3 = A_3 \; ; \; X_1 + X_2 = X_3$$

from which we deduce $X_2 = A'_2 = A_2 \cap (A_3 \ominus A_1)$.

This kind of thing still can be done in an approximate way when the A_i's are required to belong to a prescribed set of labels, such as, for instance : negative large (NL), negative medium (NM), negative small (NS), zero (0), positive small (PS), positive medium (PM), positive large (PL), unknown (?). These labels can be represented by fuzzy intervals such as the ones pictured in Figure 2. They form a (fuzzy) partition of the real line in some sense.

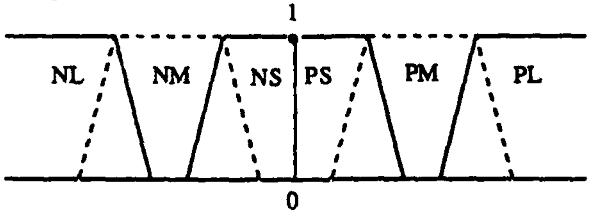


Figure 2

The condition requested to build a meaningful qualitative calculus are twofold :

C1. The advantage of qualitative reasoning is linked to the existence of symbolic calculation tables such as the ones above. Such tables should be kept when absolute orders of magnitude are introduced.

C2. The calculus, even qualitative, should remain consistent with the real line and the operations of the real line of which it is an approximation.

Standard qualitative reasoning trivially meets these requirements. However going beyond the four symbols -, 0, +, ? may look challenging. Indeed the closure property of the table seems to be incompatible with condition C2. For instance let $\mathcal{S}$ be the totally ordered set of symbols {NL, NM, NS, 0, PS, PM, PL} ; PS $\oplus$ PS = PM looks reasonable at first sight. But PS is of the form]0,a] and PS $\oplus$ PS =]0,2a] $\neq$ PM = [a,b]. Moreover $\lim_{n \to +\infty} nPS = ?$. Hence results obtained from the addition table built from $\mathcal{S}$ such that PS $\oplus$ PS = PM is inconsistent with the addition on the reals.

It does not mean that qualitative reasoning based on absolute orders of magnitude is a utopia. Interpreting orders of magnitude as intervals or fuzzy intervals apparently forbids the closure property of calculation tables. But the closure property can be preserved on <u>subsets</u> of $\mathcal{S}$ containing adjacent elements, instead of $\mathcal{S}$ itself, provided that we look for the best approximation (in the sense of inclusion) of $s_i \oplus s_j$ by means of unions of adjacent s_k's, i.e. $s_i \oplus s_j \subseteq \bigcup_{k \in K} \{s_k\}$. Note that the introduction of the symbol ? in the usual qualitative tables meets the same purpose, that is $+ \oplus - \subseteq \{-,0,+\} = ?$. What is proposed is just a generalization of the way the symbol ? appears.

The example of Figure 2 leads to consider the following term set $\mathcal{T}$ = {NL, NM, NS, 0, PS, PM, PL, [NL,NM], [NM,NS], [NS,PS]..., [NL,PM], [NM,PL], ?} where $[s_i,s_j] = \{s_k \mid s_i \leq s_k \leq s_j\}$ for $s_i \in \mathcal{S}-\{0\}$, $s_j \in \mathcal{S}-\{0\}$, $s_i < s_j$. Of course + = [PS,PL] and - = [NL,NS]. Note that if $\mathcal{S}$ has n elements distinct from 0 then $|\mathcal{T}| = (n+(n-1)+\ldots+1)+1 = \frac{(n+1)n}{2} + 1$ elements. Here $|\mathcal{T}| = 22$, for instance. This size is not so large for contemporary computers.

$\oplus$	PS	PM	PL	PM$^-$	PM$^+$	+
PS	+	PM$^+$	PL	+	PM$^+$	+
PM	PM$^+$	PM$^+$	PL	PM$^+$	PM$^+$	PM$^+$
PL	PL	PL	PL	PL	PL	PL
PM$^-$	+	PM$^+$	PL	+	PM$^+$	+
PM$^+$	PM$^+$	PM$^+$	PL	PM$^+$	PM$^+$	PM$^+$
+	+	PM$^+$	PL	+	PM$^+$	+

<u>Table 2</u> : PM$^-$ = [PS,PM] ; PM$^+$ = [PM,PL]

In Table 2 is part of the addition table (for strictly positive symbols), without any assumption regarding the model of PS, PM, PL (except that they are adjacent). Note that this Table corresponds to an associative operation, when restricted to positive values. However, it is no longer possible to preserve associativity on the whole table. This is due to the approximation procedure since associative operations remain associative when extended to intervals or fuzzy intervals. For instance with NL = -Pl, (NL $\oplus$ PM) $\oplus$ PS = $-\oplus$PS = [NL,PS], while NL $\oplus$(PM $\oplus$ PS) = NL $\oplus$ PM$^+$ = ?. However this lack of associativity does not prevent to use this approach, since the ranges which are obtained will be always valid even if they may be too large with respect to the available knowledge. Moreover, we may try to perform operations in a way where no information is lost.

The addition law can be improved (with regard to the precision of its results by subsequent requirements for instance PS ⊕ PS = PM⁻, which forces PS =]0,a], PM = [a,b] with $2a \leq b$. Note that it is not necessary to use fuzzy intervals. Adjacent intervals can do the job. However there will be discontinuity problems when the (real) values of variables cross the boundaries of the intervals modeling the symbol. Only fuzzy intervals can cope with these problems.

4 - Fuzzy relations and orders of magnitude

Orders of magnitude can be expressed in an absolute way in terms of labels such as "small", "medium" or "large" which can be represented by fuzzy intervals, as said in section 3. They can also be handled in a relative way by means of relations. This is the topic of the present section. Raiman (1985, 1986) has proposed a formal system for order of magnitude reasoning with three binary operators : Ne (for 'negligible in relation to'), Vo (for 'close to'), and Co (for 'comparable to'). Inference rules, which can be justified from a Non-Standard Analysis point of view, describe how these operators work together. See Bourgine and Raiman(1986) for an application in macroeconomics. In the following, we discuss the modeling of these operators in terms of fuzzy relations.

The idea of closeness seems to be naturally captured by an approximate equality relation. Raiman (1986) relates the ideas of closeness and of negligibility in the following way : 'x is close to y' is equivalent to '(x - y) is negligible in relation to y'. In other words, 'x is negligible in relation to y' if and only if 'x + y is close to y'. If we use an approximate equality of the form $\mu_E(x,y) = \mu_L(|x - y|)$ (as in 2.2) for modelling 'close to', the above equivalence would lead to a definition of 'negligible' which would not be relative (since $|(x + y) - y| = |x|$ does not depend on y), but absolute. It can be avoided by defining the fuzzy relation 'Vo' in terms of a quotient, i.e.

$$\mu_{Vo}(x,y) = \mu_M(\frac{x}{y}) \qquad (13)$$

where the characteristic function μ_M is such that $\mu_M(1) = 1$ and $\mu_M(t) = \mu_M(\frac{1}{t})$. Thus we have $\mu_{Vo}(x,y) = \mu_{Vo}(y,x)$ and M is a fuzzy interval which restricts values which are around 1 and which is equal to its "inverse", i.e. $M = \frac{1}{M}$ (however we have not $M^2 = 1$!). Then it leads to define the extent to which x is negligible in relation to y, by

$$\mu_{Ne}(x,y) = \mu_M(\frac{x + y}{y}) \qquad (14)$$

The combination/projection method, used in 2.1, enables us to perform the composition of Vo or of Ne with itself, or of Vo with Ne. The following results are easy to establish [2]

[2] Warning : in interval arithmetic and more generally in fuzzy arithmetic, the product MM is equal to M^2 if and only if M is either positive (i.e. $\mu_M(x) > 0 \Rightarrow x \geq 0$) or negative (i.e. $\mu_M(x) > 0 \Rightarrow x \leq 0$). Here in practice M is positive, but not (M-1).

$$\sup_{y} \min(\mu_{Vo}(x,y), \mu_{Vo}(y,z)) = \mu_{MM}(\frac{x}{z}) \geq \mu_{Vo}(x,z) \qquad (15)$$

$$\sup_{y} \min(\mu_{Ne}(x,y), \mu_{Ne}(y,z)) = \mu_{[(M-1)(M-1) \oplus 1]}(\frac{x+z}{z}) \leq \mu_{Ne}(x,z) \qquad (16)$$

$$\sup_{y} \min(\mu_{Vo}(x,y), \mu_{Ne}(y,z)) = \mu_{[M(M-1) \oplus 1]}(\frac{x+z}{z}) \geq \mu_{Ne}(x,z) \qquad (17)$$

$$\sup_{y} \min(\mu_{Vo}(x+y, z), \mu_{Ne}(y,x)) = \mu_{MM}(\frac{x}{z}) \geq \mu_{Vo}(x,z) \qquad (18)$$

They correspond to the following inference rules proposed by Raiman (1986) (for sake of brevity, here we only discuss a part of the 30 rules used in the formal system)

(i) (x Vo y) ∧ (y Vo z) → (x Vo z) ; (ii) (x Ne y) ∧ (y Ne z) → (x Ne z)

(iii) (x Vo y) ∧ (y Ne z) → (x Ne z) ; (iv) ((x + y) Vo z) ∧ (y Ne x) → (x Vo z)

The fuzzy relation approach shows that several of these rules are only "qualitatively valid". Indeed in (15), the fact that MM is a fuzzy set which contains M mirrors the intuitively satisfying lack of transitivity of the fuzzy relation Vo, strictly speaking. By contrast, as shown by (16), the relation Ne is transitive. The repeated use of the formal rules (i), (iii) or (iv) without control can lead to dubious conclusions in a way similar to sorites such as the bald man paradox (i.e., adding an hair to a bald man leaves him bald, but if we repeat the addition...). The results of the composition of fuzzy relations, such as (15)-(18), are easy to compute in terms of simple fuzzy arithmetic operations on M. The fuzzy relation calculus enables us to reason about closeness and negligibility in a rigorous way without limitations on the chaining by means of control techniques.

N.B. 1 Inference rules expressing the compatibility of the relations with respect to arithmetic operations, such as (x Vo y) ∧ (z Ne t) → xz Ne yt can be also discussed in our framework. Indeed it can be proved that

$$\sup_{\substack{x,y,z,t \\ u=xz\,;\; v=yt}} \min(\mu_{Vo}(x,y), \mu_{Ne}(z,t)) = \mu_{[M(M-1) \oplus 1]}(\frac{u+v}{v}) \geq \mu_{Ne}(u,v) \qquad (19)$$

Again we see that the rule is only "qualitatively valid", i.e. xz may be slightly less negligible with respect to yt than z in relation to t. Alternatively, we could compute what is the possibility that u is not negligible (in the sense of Ne) with respect to v, from (19).

N.B. 2 Note that we have only an approximate equality between $\mu_{Ne}(x,y)$ and $\mu_{Ne}(-x,y)$ using (14) ; a perfect equality could be recovered by modifying (14) into $\mu_{Ne}(x,y) = \mu_M(\frac{y+x}{y-x})$.

N.B. 3 Raiman (1986) makes use of a third relation Co which is such that if x Vo y, then x Co y and expresses that two values have the same sign and the same order of magnitude. We may imagine to define Co in relation to Vo and Ne in different ways, for instance by expressing that x Co y iff ∀z, x Ne z ⇔ y Ne z, following Raiman (1986). Another way would be to state that x Co y iff not[(x Ne y) ∧ (y Ne x)] in the sense of some fuzzy negation n to be chosen in relation with μ_M in order to have $\max(n[\mu_M(1 + u)], n[\mu_M(1 + \frac{1}{u})]) \geq \mu_M(u), \forall u$ (in order to guarantee $\mu_{Co} \geq \mu_{Vo}$).

5 - Concluding remarks

Other tools, not presented here, which have been also developed in fuzzy set or possibility theory, may turn to be useful in qualitative reasoning. Qualitative descriptions of the dependency between variables of the form "the more (or the less) X_1 is A_1 and... and X_n is A_n, the more (or the less) Y is B", where $A_1, \ldots, A_n$ and B are gradual properties, can be conveniently represented (by means of a special kind of fuzzy relation) and dealt with in the framework of fuzzy logic, as recently shown in Dubois and Prade (1988). Such gradual rules naturally provide a qualitative description of the behavior of systems. For instance, with $n = 2$, A_1 = 'large', A_2 = 'small', B = 'large' and the hedges "the more... the more", we express that "if X_1 increases and X_2 decreases then Y increases" (the nature of the increasingness or of the decreasingness can be modulated through a proper choice of μ_{A_1}, μ_{A_2} and μ_B).

Besides, a methodology for the control of complex dynamical systems by means of fuzzy expert rules which provide a qualitative description in terms of fuzzy sets of the relation between action variables and observable state variables, was settled more than ten years ago (Mamdani and Assilian, 1975) ; see Sugeno (1985) for an overview of existing applications. People in Artificial Intelligence have also considered the problem of qualitative control recently (e.g. Clocksin et Morgan, 1986).

The intended purpose of this short communication is to point out that fuzzy set and possibility theory can offer valuable tools for qualitative reasoning problems. In particular "commonsense" arithmetic reasoning (e.g. Simmons, 1986) can be easily handled using fuzzy intervals and fuzzy comparison relations. This framework is especially useful for interfacing numerical data and symbolic information.

References

d'Ambrosio B. (1987) Extending the mathematics in qualitative process theory. Proc.6th National Conf. on Artificial Intelligence (AAAI-87), Seattle, July 13-17, 595-599.

Bellman R., Zadeh L.A. (1977) Local and fuzzy logics. In : Modern Uses of Multiple-Valued Logic (J.M. Dunn, G. Epstein, eds.), Reidel Publ., Dordrecht, 103-165.

Bobrow D.G. (ed.) (1984) Qualitative Reasoning about Physical Systems (with papers by J. De Kleer, K.D. Forbus, B. Kuipers, B.C. Williams,...). North-Holland, Amsterdam. Also special volume of Artificial Intelligence, 24, 1984.

Bourgine P., Raiman O. (1986) Economics as reasoning on a qualitative model. Proc. Inter. Conf. on Economics and Artificial Intelligence, Aix-en-Provence, September, 185-189.

Clocksin W.F., Morgan A.J. (1986) Qualitative control. Proc. 7th European Conf. on Artificial Intelligence, Brighton, July, Vol. 1, 350-356.

Davis E. (1987) Constraint propagation with interval labels. Artificial Intelligence, 32, 281-331.

Dormoy J.L. (1987) Résolution qualitative : complétude, interprétation physique et contrôle. Mise en œuvre dans un langage à base de règles : BOOJUM. Doct. Thesis, Univ. Paris VI, December.

Dubois D. (1987) Linear programming with fuzzy data. In : Analysis of Fuzzy Information, Vol. 3 : Applications in Engineering and Sciences (J.C. Bezdek, ed.), CRC Press, Boca Raton, Fl., 241-263.

Dubois D., Prade H. (1980) Fuzzy Sets and Systems : Theory and Applications. Academic Press, New York.

Dubois D., Prade H. (1985) (with the collaboration of Farreny H., Martin-Clouaire R., Testemale C.) Théorie des Possibilités. Applications à la Représentation des Connaissances en Informatique. Masson, Paris, (2nd revised and extended edition, 1987). English version : Possibility Theory. An Approach to Computerized Processing of Uncertainty. Plenum Press, New York, 1988.

Dubois D., Prade H. (1987) Fuzzy numbers : an overview. In : Analysis of Fuzzy Information - Vol. 1 : Mathematics and Logic (J.C. Bezdek, ed.), CRC Press, Boca Raton, Fl., 3-39.

Dubois D., Prade H. (1988) Gradual inference rules in approximate reasoning. Submitted.

Mamdani E.H., Assilian S. (1975) An experiment in linguistic synthesis with a fuzzy logic controller. Inter. J. Man- Machine Studies, 7, 1-13.

Raiman O. (1985) Raisonnement qualitatif. Tech. Rep. n° F093, Centre Scientifique IBM France, Paris, November.

Raiman O. (1986) Order of magnitude reasoning. Proc. 5th National Conf. on Artificial Intelligence (AAAI-86), Philadelphia, PA, August, 100-104.

Simmons R. (1986) "Commonsense" arithmetic reasoning. Proc. 5th National Conf. on Artificial Intelligence (AAAI-86), Philadelphia, PA, August, 118-124.

Sugeno M. (ed.) (1985) Industrial Applications of Fuzzy Control. North-Holland, Amsterdam.

Travé L., Kaszkurewicz E. (1986) Qualitative controllability and observability of linear dynamical systems. Proc. IFAC Cong. Large Scale Systems, Zürich, Switzerland, August, 964-970.

Zadeh L.A. (1979) A theory of approximate reasoning. In : Machine Intelligence, Vol. 9 (J.E. Hayes, D. Michie, L.I. Mikulich, eds.), Elsevier, N.Y., 149-194.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique